БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ

ИНСТИТУТ ПО ИНФОРМАЦИОННИ И КОМУНИКАЦИОННИ ТЕХНОЛОГИИ

Атанас Петров Узунов

ДЕТЕКЦИЯ НА ГОВОР В СИСТЕМИ ЗА РАЗПОЗНАВАНЕ НА ДИКТОРИ

Д И С Е Р Т А Ц И Я

за присъждане на образователна и научна степен

“ДОКТОР“

По направление:
4.6. Информатика и компютърни науки
Научна специалност: 01.01.12. Информатика

Научен консултант:

доц. д-р Георги Глухчев

София
2020

Благодарности

Бих искал да изкажа искрената си благодарност на моя консултант доц. д-р Георги Глухчев за полезната дискусия и напътствия по време на подготовката на дисертационния ми труд. Бих искал също така да благодаря на доц. д-р Божан Жечев за направените коментари и предложения.

Бих искал да благодаря и на семейството си, защото без тяхната вяра и подкрепа този проект никога нямаше да бъде завършен.

Списък на използваните основни съкращения[1]

HMM – *Hidden Markov Model* – скрит Марковски модел;
GMM – *Gaussian mixture model* – модел използващ смес от Гаусови разпределения;
DTW – *Dynamic Time Warping* – метод за сравнение на две времеви последователности чрез динамично изменение на оста на времето;
VAD – *Voice Activity Detection* – детекция на говор;
ED – *Endpoint Detection* – определяне на граничните точки на говорно съобщение;
LTSD – *Long-Term Spectral Divergence* – дълговременна спектрална дивергенция;
DFT – *Discrete Fourier Transform* – дискретна трансформация на Фурие;
FFT – *Fast Fourier Transform* – бърза трансформация на Фурие;
DCT – *Discrete Cosines Transform* – дискретна косинус трансформация;
DWT – *Discrete Wavelet Transform* – дискретна уейвлет трансформация;
FDLP - *Frequency Domain Linear Prediction* - линейно предсказване в честотната област;
LTSV - *Long-Term Signal Variability* - дълговременна изменчивост на сигнала;
LSFM - *Long-term Spectral Flatness Measure* - дълговременна гладкост на спектъра;
UBM - *Universal Background Model* - универсален фонов модел;
SVM –*Support Vector Machine* – метод на опорните вектори;
RBF – *Radial Basis Function* – радиална базисна функция;
DET - *Detection Error Tradeoff* – графика визуализираща изменението на грешката при бинарна класификация;
PLDA - *Probabilistic Linear Discriminant Analysis* - вероятностен линеен дискриминантен анализ;
NIST - *National Institute of Standards and Technology* - Национален институт по стандарти и технология в САЩ;
SRE - *Speaker Recognition Evaluations* – оценъчни сесии в областта на разпознаване на диктори организирани от NIST;
SNR – *Signal-to-Noise Ratio* – отношение сигнал/шум*;*
ROC curve - *Receiver Operating Characteristics curve* – специфичен термин, с който се означава графика позволяваща да се оцени качеството на бинарна класификация;
SGMM - *Sequential Gaussian Mixture Model* - модел с Гаусови смеси с последователна адаптация;
VQ – *Vector Quantization* – векторно квантуване;
LLR – *Log-Likelihood Ratio* – логаритмично отношение на правдоподобие;
JFA - *Joint Factor Analysis* – съвместен факторен анализ;
MLP – *Multi-Layer Perceptron* - многослоен перцептрон – клас невронни мрежи;
DNN - *Deep Neural Network* - невронна мрежа с дълбочинно обучение;
RSF - *Running Spectral Filters* - метода с изместващи се спектрални филтри;
CMS – *Cepstral Mean Subtraction* – метод за центриране на кепстрална последователност;

[1] Всички английски термини и съкращения са обяснени подробно в текста при първото им появяване.

Увод

Биометриката е наука за разпознаване на личността на човек чрез анализ посредством технически средства на неговите физически или поведенчески характеристики. Тя се основава на предположението, че много от тези характеристики (модалности) са строго индивидуални. Имат се предвид следните физически характеристики: глас, лице, ирис, отпечатъци на пръстите, геометрия и вени на ръката, форма на ухото и съответно поведенчески такива като подпис, ръкописен стил, динамика на писане на клавиатура, походка и др. [Kisku et al., 2014].

През последното десетилетие се наблюдава експлозивно развитие (в САЩ и в Китай) на биометричните технологии и тяхното приложение в различни области – от хранителни магазини, летища до правителствени учреждения. Необходимостта от биометрични решения насочва огромни инвестиции в изследвания, което води до разработка на нови алгоритми за извличане на признаци и класификация и на авангардни приложения.

Гласът като една от основните модалности е най-достъпният биометричен признак. Това е така поради масовото разпространение в последните години на мобилни телефони и приложения за пренос на глас по интернет (VoIP). Подобна масовост на техническите средства за пренос на глас води до разработването на значително повече приложения в областта на гласовата биометрика, отколкото такива за другите модалности.

Понастоящем приложенията в областта на гласовата биометрика могат да бъдат разделени в три основни групи [Jain et al., 2008]:

- speaker detection (speaker spotting) – сепариране на глас (диктор) чрез анализ на множество разговори (например в кол-центрове);
- speaker verification (voice authentication) – удостоверяване автентичността на даден глас – типично приложение е дистанционен контрол на достъпа (например банкови транзакции);
- forensic speaker recognition - разпознаване на диктори в криминалистиката;

Много бързо развиващо се направление е гласовата биометрика за мобилни устройства. Характерно за този вид мобилна биометрика е фактът, че разговорите чрез мобилни устройства обикновено се реализират в изключително динамична среда.

В действителност изброените по-горе приложения винаги се базират на система (локална или дистанционна), която решава задачата за разпознаване на диктори.

Независимо каква ще бъде задачата – зависимо или независимо от текста разпознаване на диктори, верификация или идентификация, тази система трябва да включва един задължителен алгоритъм (модул), а именно детектор на говор. Той локализира говорните фрагменти в постъпилия в системата аудио поток и предава информацията за тях за по-нататъшна обработка. Реално неговото функциониране е от ключово значение за цялата система. Това е така, защото при създаване на модела на гласа на диктора се използват само говорни фрагменти и точността, с която те се локализират оказва съществено влияние върху крайното решение на биометричната система.

Експлозивното развитие в световен мащаб на биометричните технологии респективно на гласовата биометрика определя задачата за разпознаване на личността като изключително актуална, от който факт следва и актуалността на проблемите свързани с детекцията на говор и разглеждани в дисертационния труд.

Дисертационния труд се състой от пет глави. Дисертацията съдържа 164 страници, 48 фигури, 27 таблици, 151 цитирани източника.

Първа Глава е озаглавена ***„Детекция на говор. Аналитичен преглед.“***. В нея е направен подробен аналитичен преглед на алгоритмите за детекция на говор в контекста на разработените в последното десетилетие системи за разпознаване на диктори. Описани са двата вида детектори на говор – детектори за определяне на гранични точки на кратки фрази и детектори на говорни фрагменти (VAD-алгоритми). В общия случай подобни детектори са реализирани като класификатори и тук поотделно са разгледани двата основни етапа при тях – избор на признаци и класификационни правила. В обзора е отделено основно внимание на системи работещи със сертифицирани корпуси с говор записан по телефонен канал – NIST, CTS, Switchboard и др. Акцентът в изложението е върху характеристиките на признаците, които са използвани в различните детектори. Разгледани са подробно някой от най-значимите понастоящем системи за разпознаване на диктори и е анализирана ролята на детекторите на говор в тях. Посочена е успешната стратегия при разработка на алгоритми за детекция на говор в реална среда, а именно комбинацията на източници, които доставят различна информация. Тази стратегия включва комбинация на различни свойства на говорния сигнал в един признак, на различни признаци в един VAD-алгоритъм и на различни VAD-алгоритми работещи съвместно. От своя страна съвместно работещите VAD-алгоритми могат да бъдат изградени с различни класификатори, което дава възможност за по-голяма адаптивност на детектора на говор при промяна в условията на средата.

Втора Глава е озаглавена ***„Дефиниране на признаци за детекция на говор използващи свойствата на САКФ и СГЗ.“***. В нея са разгледани някои характеристики на спектралната автокорелационна функция, получена чрез спектъра на Фурие. Предложен е метод, при който чрез прилагане на делта-филтър върху спектралната автокорелационна функция е получена т.н. делта спектрална автокорелационна функция. Анализирани са особеностите на тази филтрация, при която е постигнато усилване в честотната област на хармоничната структура на говорния сигнал. Предложени са два подхода за изчисляване на признаци за детекция на говор. При първия подход признаците се определят само чрез свойствата на делта спектралната автокорелационна функция. По този начин са дефинирани три признака като първия от тях (т.н. MD-признак) е в скаларна форма и е предназначен за детекция чрез анализ на времеви контури, докато другите два (т.н. BMD- и MMD-признаци) са вектори и са предназначени за детекция чрез алгоритми за разпознаване. При втория подход признаците се получават чрез комбиниране на свойствата на делта спектралната автокорелационна функция и тези на модифицирания спектър на групово закъснение. По този начин са дефинирани два признака (т.н. lin-GDMD и log-GDMD - признаци) които са в скаларна форма и са предназначени за детекция на говор чрез анализ на времеви контури. В тази глава са разгледани и основните методи за определяне на спектъра на групово закъснение и е извършен теоретичен анализ на изменението му при зашумени с адитивен шум говорни сигнали. Анализът е реализиран косвено, чрез изследване на изменението на аргументите на проекционните функции на сходство на основата на адитивния спектрален модел.

Трета Глава е озаглавена ***„Алгоритми за определяне на гранични точки при зависима от текста верификация на диктори. Експериментално изследване.“.*** В нея е предложен подход за определяне на гранични точки на говорно съобщение включващ алгоритъм за изчисляване на адаптивни прагови стойности и детерминиран краен автомат. На базата на този подход са разработени три алгоритъма за определяне на гранични точки, които са формирани съобразно използваните времеви контури. Оценката на ефективността на предложените алгоритми е реализирана на два етапа. На първия етап е оценена тяхната точност на детекция чрез хистограмен анализ на разликите между ръчно определените и получените от съответния алгоритъм гранични точки. Експериментите са реализирани със зашумени говорни данни на български (от корпус BG-SRDat) и английски език (от корпус TIDIGITS). На втория етап е оценено влиянието на предложените алгоритми за определяне на гранични точки върху грешката при

разпознаване в две системи за зависима от текста верификация на диктори базирани съответно на DTW и HMM алгоритми. При експериментите са използвани говорни данни на български език записани по телефонен канал (от корпус BG-SRDat).

Четвърта Глава е озаглавена *„Алгоритми за детекция на говор при независима от текста идентификация на диктори. Експериментално изследване.“*. В нея за предложени два алгоритъма за детекция на говор условно означени като VAD-1 и VAD-2. При VAD-1 се използва невронен класификатор с многослоен перцептрон и признаци във векторна форма. При VAD-2 се използват признаци в скаларна форма и прагова логика. Оценката на ефективността на алгоритмите за всеки признак (референтен и предложен от автора в Глава 2) е получена на два етапа. На първия етап е оценена точността на детекция. Тя е определена чрез анализ на разликите между ръчно и автоматично (от съответния алгоритъм) локализирани говорни сегменти. Изчислени са няколко вида грешки всяка от които описва различни характеристики на VAD алгоритмите. Допълнително са изчислени и параметри, които служат за оценка на точността на бинарната класификация. На този етап са реализирани експерименти със зашумени говорни данни на български (от корпус BG-SRDat) и английски език (от корпуси NOIZEUS, TIDIGITS, TIMIT). На втория етап е оценено влиянието на алгоритмите за детекция на говор върху точността на разпознаване. За всеки признак (референтен и предложен от автора) се формира отделен детектор на говор (VAD-1 или VAD-2), който става част от система за независима от текста идентификация на диктори реализирана чрез невронна мрежа (многослоен перцептрон). Тук експериментите са реализирани с говорни данни на български език записани по телефонен канал (от корпус BG-SRDat).

Пета Глава е озаглавена **„*BG-SRDat – Корпус с говорни данни записани по телефонен канал и предназначен за разпознаване на диктори*“**. В нея е описан корпуса BG-SRDat (Bulgarian language Speaker Recognition DATa) съдържащ говор записан по телефонен канал (стационарни и мобилни телефони и чрез VoIP) и включващ фрази и разговори на български и само фрази на английски език. Акцента при изграждане на корпуса е многообразието на комуникационната среда, т.е. различни телефонни канали, различно местоположение на диктора, различен съпътстващ шум при произнасяне на фразите и др. Корпуса съдържа 630 записа с различна продължителност, събрани от 40 диктора-мъже. Експерименталните изследвания описани в представения дисертационен труд са реализирани основно с говорни данни избрани от този корпус.

ГЛАВА 1

Детекция на говор. Аналитичен преглед.

По същество всички модели са грешни,
но някои са полезни.
George Box, математик, 1987

Не вярвам в наука,
която не е свързана с практиката.
Не вярвам в образование,
което не е свързано с практиката и науката.
Не вярвам в бизнес,
който не е свързан с образованието и науката.
Hermann Graf, банкер, 2016

1.1. Увод

Детекцията на говор се дефинира като процес на локализация на говор на фона на различни видове не-говорни събития. Под не-говорни събития се разбират всички звукови събития съпътстващи реализацията на говорното съобщение, но не свързани с информацията, която то пренася. Тези не-говорни събития може да са от заобикалящата среда (уличен шум, странични разговори и др.), от комуникационния канал или да са звукови артефакти, генерирани от диктора (въздишка, кашлица и др.).

Работата на биометричните системи за верификация и идентификация на диктори силно зависи от качеството на сегментация на говорната последователност. Тъй като обучението на модела на диктора се основава изцяло на говорни фрагменти, то точността на алгоритъма, който локализира тези фрагменти влияе съществено върху точността на разпознаване.

Детекцията на говор е означена в англоезичната литература с термините Speech/non-speech detection, Voice activity detection и Speech activity detection, за които в повечето литературни източници се приема, че са синоними. От тях най-разпространен е Voice Activity Detection (VAD) [Tuononen, 2008]. Като подзадача на детекцията на говор, а понякога и като отделен вид детекция се разглежда т.н. Определяне на Граничните Точки (ОГТ) на говорното съобщение. При него се локализират само

граничните точки (начална и крайна) на съобщението докато паузите вътре в думата или фразата не се маркират (ако са до определена дължина). ОГТ в англоезичната литература е означен като Endpoint Detection (ED). В повечето случаи ED-алгоритмите се използват при зависимо от текста разпознаване на диктори, където се работи с кратки думи или фрази.

Терминологично по-точно би било да се приеме, че детекцията на говор (Speech Detection) има две подзадачи – откриване на говорни сегменти в даден аудио сигнал т.е. Voice Activity Detection (бинарна класификация) и определяне на гранични точки – Endpoint Detection. По-универсално значение при детекцията на говор имат VAD-алгоритмите и те намират приложение предимно при независимо от текста разпознаване на диктори. Освен това по локализираните от тях говорни фрагменти могат да се определят граничните точки на говорните данни, без да е необходимо да се разработва специален алгоритъм за тази цел. Най-често ED-алгоритмите се използват в задачи, при които са налице съществени ограничения от гледна точка на време за изпълнение и изчислителни ресурси. Тъй като съществуват значителни различия при реализациите на двата вида алгоритми - ED и VAD - то в представения обзор те ще бъдат разгледани в отделни параграфи.

Детектора на говор е отделен етап в предварителната обработка на биометричната система. Основната цел при разработването на този вид алгоритми е да се постигне робастност на тяхното решение, т.е. сегментацията на говорната последователност да не се променя независимо от промяната на качеството на сигнала и условията на средата. Подобна стабилност на детектора на говор е изключително важна за надеждната работа на биометричната система в реална среда (unconstrained environments) например като мобилно банкиране използващо разпознаване на глас или автоматично разпознаване на диктори за целите за сигурността [Nautsch et al., 2016].

Алгоритми за детекция на говор се разработват от няколко десетилетия в редица области като подобряване качеството на речта (speech enhancement), разделяне (сепариране) на диктори (speaker diarization), разпознаване на говор, разпознаване на диктори, кодиране на говор и др. Актуално направление в последните години е детекция на синтетична реч (synthetic speech), получена чрез различни методи за преобразуване на глас и синтез на говор. Тази реч се използва за преодоляване (spoofing attacks) на гласовата биометрична защита на системите за сигурност [Sahidullah et al., 2015], [Das et al., 2019], [Sailor et al., 2018]. Интересно направление е и детекцията на говор произнесен като шепот [Ashihara et al., 2019].

Необходимо е да се посочат три обстоятелства. Първо, всяка една от областите налага специфични изисквания при разработка на съответния детектор. Второ, редица от публикуваните алгоритми не са подложени на стандартни тестове включващи едни и същи данни, шум, комуникационна среда и др. което от своя страна води до затруднения при сравняване на тяхната ефективност [Kola et al., 2011]. Трето, има публикации, в които е направена оценка единствено на точността, с която се локализират говорните фрагменти, т.е. разгледаните детектори не са предназначени за конкретна област или приложение.

Поради посочените обстоятелства в този обзор интерес представляват публикации предимно от последното десетилетие, в които детектора на говор е част от система за разпознаване на диктори и при реализираните експерименти са използвани сертифицирани корпуси с говорни данни. Освен това се имат предвид алгоритми за детекция използващи само един източник на говорен сигнал [Ramirez et al., 2007]. Проблемите на компенсацията на шума и подобряване качеството на речта намират място в редица алгоритми за детекция говор, но те са извън темата на представения обзор (и дисертационен труд) и ще бъдат отбелязани, където е необходимо, но няма да бъдат коментирани.

При оформянето на обзор е възможно да се използват два подхода. При първия подход алгоритмите за детекция на говор в анализираните литературни източници се разделят в няколко групи, например: традиционни, статистически и от областта на разпознаване на образи. Разглеждат се само основните параметри на алгоритмите и как те са реализирани в дадено изследване. В този случай обаче се изпускат редица детайли описани в конкретните реализации. При втория подход, автора избира няколко публикации, в които са представени значими резултати и ги описва в цялост. Тук освен общата информация към коя група принадлежи даден алгоритъм са налице и редица полезни за читателя детайли от разглежданото изследване. Обзора в представения дисертационен труд ще бъде реализиран чрез използване на втория подход.

1.2. VAD-алгоритми

Детектора на говор съдържа три основни модула: извличане на признаци, приемане на решение и допълнителна корекция (hangover scheme) [Ramirez et al., 2007].

Често използваните признаци се основават на: спектрална дивергенция [Ramirez et al., 2004], функции на групово закъснение [Krishnan et al., 2006], автокорелационни функции [Ghaemmaghami et al., 2010a], периодични и апериодични компоненти [Ishizuka

et al., 2010], делта-фазов спектър [McCowan, 2012], форманти [Yoo et al., 2015], полиномиална регресия на Мел-спектъра [Disken, 2017], i-вектори [Yamamoto et al., 2017].

Модула за приемане на решение (класификатор) използва различни подходи съобразно решаваната задача и вида на корпуса с говорни данни. Например при зависимо от текста разпознаване на диктори с корпуса RSR2015 [Alam et al., 2014] са използвани класификатори със самообучение на базата на: векторно квантуване [Kinnunen et al., 2013], модел с Гаусови смеси и модел с Гаусови смеси с последователна адаптация (sequential Gaussian mixture model) [Ying et al., 2011]. При независима от текста верификация на диктори в NIST 2008 SRE (Speaker Recognition Evaluation) е използван многослоен перцептрон [Ganapathy et al., 2011], а в NIST 2010 SRE е намерил приложение модела на Гаусови смеси с обучение [Mak et al., 2014]. При същия тип верификация на диктори и NIST 2016 SRE е използвана невронна мрежа с дълбочинно обучение (deep neural network - DNN) [Yamamoto et al., 2017].

1.2.1. Признаци използвани при VAD- и ED-алгоритмите

В текста ще бъдат описани редица признаци, използвани при VAD- и ED-алгоритмите. В основната си част материала в тази подточка се базира на обзора публикуван в [Graf et al., 2015].

1.2.1.1. Мощност (енергия)на сигнала. Спектър.

Мощността (респективно енергията) на сигнала е един от първите признаци, използван при детекция на говор. Обикновено високите стойности при този признак означават наличие на говор, а ниските – за наличие на шум или пауза. Това свойство е в основата на функционирането на различните детектори анализиращи времевия контур на мощността или на енергията и използващи праг. Основния проблем в случая е как да бъде изчислена праговата стойност като се има предвид и нивото на шума.

В [Ramirez et al., 2004] е предложен алгоритъма анализиращ спектрална информация в границите на няколко сегмента. Ако амплитудния спектър за *n-тия* сегмент има вида $|X(n,k)|$ където $k = 0,...,K-1$ и K е размера на трансформацията на Фурие, то дълговременната обвивка на спектъра (Long-Term Spectral Envelope - LTSE) от *M-ти* ред има вида

$$LTSE_M(n,k) = \max\left\{\left|X(n+j,k)\right|\right\}_{j=-M}^{j=+M} \quad (1.1)$$

За *n*-тия сегмент дълговременната спектрална дивергенция (Long-Term Spectral Divergence - LTSD) от *M-ти* ред между спектрите на говора и шума е дефинирана като отклонение на LTSE спрямо средния амплитуден спектър на шума $|S(k)|$ и има вида

$$LTSD_M(n) = 10\log_{10}\left(\frac{1}{K}\sum_{k=0}^{K-1}\frac{LTSE_M^2(n,k)}{|S(k)|^2}\right), \quad (1.2)$$

В работата текущата стойност на LTSD се сравнява с адаптивен праг и ако тя е по-голяма се приема, че имаме говорен сегмент, а ако не, то сегментът е шум и в този случай се извършва адаптация на спектъра на шума. Освен това се предполага, че първите сегменти на анализирания файл съдържат само шум и чрез тях се получава началната оценка на неговия спектър.

1.2.1.2. Основен тон и хармонична структура на спектъра

Известно е, че всички гласни и дори някои съгласни имат хармонична структура, която е основополагаща характеристика на говорния сигнал. Признаци, които описват по подходящ начин тази структура се считат за надеждни индикатори за наличието на говор. Подобни признаци не са ефективни само в два случая. Първо, когато говорните фрагменти съдържат фрикативни съгласни и второ, когато в анализирания сигнал е налице музика или шум с хармонична структура [Graf et al., 2015].

В [Kristjansson et al., 2005] са разгледани признаци за детекция на говор, които се основават на хармоничната структура на говорния сигнал. Във времевата област те са: брой пикове в автокорелационната функция (АКФ), максимален пик на АКФ, максимален пик на АКФ получена чрез метода на линейно предсказване (Maximum LPC Residual Autocorrelation Peak) и енергия на АКФ (Windowed Autocorrelation Lag Energy). В спектралната (кепстрална) област са спектрална ентропия, спектрална автокорелационна функция (Spectral Autocorrelation Peak Valley Ratio - SAPVR) и максимума в кепстралните коефициенти (Cepstrum Peak).

АКФ във времевата област е често използван признак при детекция на говор, главно заради възможността за откриване на периодичност в анализирания сигнал. Кратковременната нормализирана АКФ има вида

$$acorr_m(k) = \frac{\sum_{n=k}^{N} x_m(n)x_m(n-k)}{\left(\sum_{n=1}^{N-k} x_m(n)^2\right)^{1/2}\left(\sum_{n=k}^{N} x_m(n)^2\right)^{1/2}} \quad (1.3)$$

където x_m е *m-я* сегмент и *k* е изместването (лаг) на АКФ.

Описанието на изброените по-горе на признаци е както следва:

- *Максимален пик на АКФ.* Дефинира се като амплитудата на максималния пик на АКФ в диапазона от изместванията (лагове) съответстващ на диапазона на изменение на основния тон.
- *Енергия на АКФ*. Въведен е преместващ се прозорец в автокорелационната област. За всяко преместване се изчислява енергията на АКФ в границите на прозореца и се определя максималната й стойност сред всички изместванията, или

$$WALE(n) = \max_{l} \sum_{i=l}^{l+W-1} \left| acorr_n(i) \right|^2 \tag{1.4}$$

  където $accor_n$ е векторът на АКФ, W е дължината на прозореца в автокорелационната област. Допълнително е дефиниран признак multi-frame WALE ($WALE_{MF}$) във вида

$$WALE_{MF}(n) = \max_{l} \sum_{i=l}^{l+W-1} \sum_{t=n-\alpha}^{n+\beta} \left| acorr_t(i) \right|^2 \tag{1.5}$$

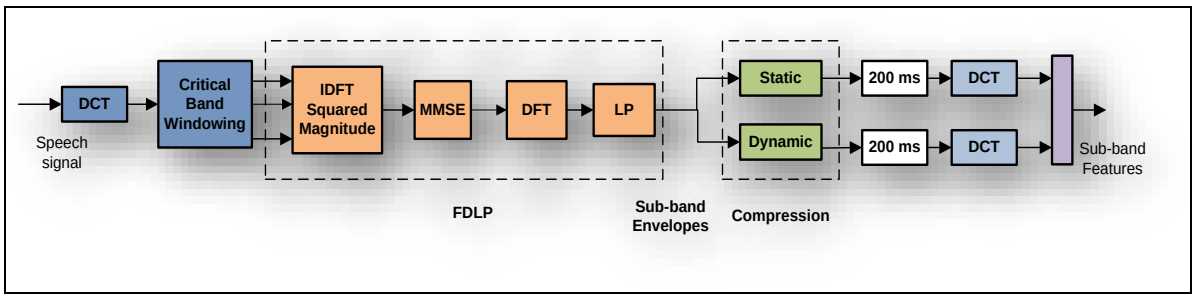
  където α и β показват колко предишни и бъдещи сегмента ще бъдат включени в изчисленията.
- С*пектрална ентропия.* Кратковременния спектър на говорния сигнал се разглежда като вероятностно разпределение на дискретна случайна величина и ентропията се изчислява за това разпределение. Предполага се, че хармоничната структура на гласните ще доведе до сравнително ниски стойности на ентропията, а фоновия шум – съответно до високи.
- *Спектрална автокорелационна функция (САКФ).* Тук е използван амплитудния спектър, получен чрез преобразуване на Фурие. При звучни фрагменти в САКФ се наблюдават регулярно разпределени пикови стойности. Признака се изчислява като отношение между сумата от пиковете и стойността на първия локален минимум (first valley) в САКФ.
- *Кепстрален максимум.* Този признак се основава на свойството на кепстъра да трансформира произведението в спектралната област на предавателните характеристики на гласовия източник и гласовия тракт в сума от кепстрални компоненти. Бързо променящия се спектър на гласовия източник се представя в областта на кепстралните коефициенти с висок пореден номер, т.е. хармоничните компоненти се изявяват чрез пик в тази област. Предложения признак се изчислява като разлика между максималната и минималната стойности в кепстъра.

1.2.1.3. Спектър на модулациите

Спектъра на времевите траектории на спектралните обвивки на говорния сигнал се нарича спектър на модулациите на говора (modulation spectrum of speech). Този спектър при непрекъсната реч е доминиран от честотни компоненти между 2 Hz и 8 Hz [Ganapathy et al., 2010].

Един от алгоритмите за определяне на спектъра на модулациите е чрез използване на линейно предсказване в честотната област (Frequency Domain Linear Prediction - FDLP). За да се използва алгоритъма за FDLP е необходимо предварително да се извърши следната обработка [Ganapathy et al., 2010]. Върху входния сигнал се прилага DCT. Така получения в честотната област сигнал се умножава с тегловни функции разположени в Bark-скала, за да се получат лентовите (sub-band) DCT компоненти. Тези компоненти са входни данни за FDLP алгоритъма. За всяка лента се прилага обратно преобразуване на Фурие върху DCT коефициентите, за да се получи аналитичен сигнал. Чрез прилагане на преобразуване на Фурие върху квадрата на амплитудата на аналитичния сигнал се получава спектралната автокорелационна функция. Чрез линейно предсказване, приложено върху тази функция се получават времевите обвивки (контури) на Hilbert в отделните ленти (sub-band temporal envelopes) и те са изходните данни на FDLP алгоритъма.

Тези времеви контури се преобразуват чрез статичен и динамичен компресиращи алгоритми. Чрез прилагане на дискретно косинус преобразуване върху компресираните последователности се получават съответно статичния и динамичния спектър на модулациите [Ganapathy et al., 2010]. На фиг. 1.1 е показана блок схемата за определяне на признаци на базата на спектъра на модулациите [Ganapathy et al., 2011].



Фиг. 1.1. Блок схема за определяне на признаци на базата на спектъра на модулациите (Фигурата е адаптирана версия на фиг. 1 в [Ganapathy et al., 2011]).

1.2.1.4. Стационарност

Измененията във времето на шума обикновено са значително по-бавни от тези на говорния сигнал. Ако шума се разглежда като стационарен сигнал то степента на

нестационарност може да използва при детекция на говор. В [Ghosh et al., 2011] е предложена т.н. дълговременна изменчивост на сигнала (Long-term signal variability – LTSV) като мярка за нестационарност. Тя се дефинира за *m*-тия сегмент като дисперсия на ентропията за всички дискретни честоти съгласно

$$LTSV(m) = \frac{1}{K}\sum_{k=0}^{K-1}\left(E(m,k) - \frac{1}{K}\sum_{k=0}^{K-1} E(m,k) \right)^2 \tag{1.6}$$

където ентропията $E(m,k)$ е изчислена за *R* предходни сегмента спрямо *m*-тия сегмент и се дефинира като

$$E(m,k) = -\sum_{n=m-R+1}^{m} \frac{S(n,k)}{\sum_{l=m-R+1}^{m} S(l,k)} \log\left(\frac{S(n,k)}{\sum_{l=m-R+1}^{m} S(l,k)} \right) \tag{1.7}$$

и $S(n,k) = \left|X(n,k)\right|^2$ и $\left|X(n,k)\right|$ е амплитудният спектър за *n*-тия сегмент, $k = 0,...,K-1$, K е размерът на трансформацията на Фурие.

Дефинираната в (1.7) ентропия отразява времевата гладкост на спектъра за всяка дискретна честота в границите на времеви прозорец с дължина *R* сегмента. При стационарни сигнали спектъра не се променя във времето и тази ентропия ще има максимална стойност. За стационарни сигнали LTSV клони към нула. Когато е налице говор, в някои честотни ленти се появяват значителни изменения в ентропията, което води до големи стойности на дисперсията.

Тази идея е доразвита в [Tsiartas et al., 2013] като дисперсията не се определя за всички честотни ленти, а се изчисляват отделни дисперсии за различни честотни диапазони. По този начин се намалява влиянието на нестационарни шумови компоненти.

В [Ma et al., 2013] е предложена дълговременна мярка за гладкостта на спектъра (Long-term spectral flatness measure - LSFM). При този признак вместо ентропията е използвано отношението между геометрична и аритметична средни стойности на спектъра, получени в границите на *R* времеви сегмента във вида

$$LSFM(n) = \frac{1}{K}\sum_{k=0}^{K-1} \log\left(\frac{(\prod_{r=0}^{R-1} X(k,n-r))^{1/R}}{(\sum_{r=0}^{R-1} X(k,n-r))/R} \right) \tag{1.8}$$

$$\Lambda(Y) = L(Y \mid \lambda_{speech}) - L(Y \mid \lambda_{noise}) \tag{1.9}$$

Тъй като стойността на LSFM е винаги отрицателна, то наличието на говор се установява при по-голяма абсолютна стойност на признака.

1.2.2. Класификатори, използвани при VAD- и ED-алгоритмите

В текста ще бъдат разгледани някой основни класификатори намерили приложение при VAD-алгоритмите.

1.2.2.1. Модел с Гаусови смеси

GMM е генеративен модел използван доста често при VAD-алгоритмите [Ferrer et al., 2013], [Hanilci et al., 2015]. При него всеки клас се представя като претеглена сума от *M* Гаусови разпределения

$$p(x \mid \lambda) = \sum_{i=1}^{M} \omega_i p_i(x), \tag{1.10}$$

където ω_i е тегловния коефициент за отделното Гаусово разпределение и $p_i(x)$ е *D*-размерно Гаусово разпределение със среден вектор μ_i и ковариационна матрица Σ_i. Обозначението на модела е $\lambda = \{w_i, \mu_i, \Sigma_i\}_{i=1}^{M}$. Параметрите на всеки клас се определят поотделно чрез Expectation-maximization (EM) - алгоритъма на базата на критерия на максималното правдоподобие (Maximum-likelihood criterion). При тест, имайки двата модела на говора λ_{speech} и на шума λ_{noise} и тестовия вектор $Y = \{y_1, ..., y_T\}$ то разстоянието между тях се определя като

$$\Lambda(Y) = L(Y \mid \lambda_{speech}) - L(Y \mid \lambda_{noise}) \tag{1.11}$$

където

$$L(Y \mid \lambda) = \frac{1}{T} \sum_{t=1}^{T} \log p(y_t \mid \lambda) \tag{1.12}$$

е средната стойност на логаритмичната функция на правдоподобие (log-likelihood) на *Y* при даден GMM λ.

При GMM се използва т.н. универсален фонов модел (Universal Background Model - UBM) който се обучава с възможно максимален брой диктори и представлява генералната съвкупност на векторите на признаците. Моделите λ_{speech} и λ_{noise} се получават чрез адаптация на UBM с помощта на метода използващ максимума на апостериорната вероятност (maximum a posteriori adaptation – MAP-адаптация). Средните вектори $\hat{\mu}_i$ за даден модел се получават чрез адаптация във вида

$$\begin{aligned} \hat{\mu}_i &= \alpha_i E_i(x) + (1 - \alpha_i) \mu_i^{UBM} \\ \alpha_i &= n_i / (n_i + r) \end{aligned} \tag{1.13}$$

Където α_i е коефициент на адаптация, n_i (probabilistic count) и $E_i(x)$ са достатъчни статистики от нулев и първи ред във вида

$$n_i = \sum_{t=1}^{T} p_t(i \mid x_t)$$
$$E_i(x) = \sum_{t=1}^{T} p_t(i \mid x_t) x_t \tag{1.14}$$

където x_t е *t*-я параметричен вектор от анализираната фраза и $p_t(i \mid x_t)$ е апостериорната вероятност *t*-я вектор да принадлежи на *i*-я (mixture component) компонент на UBM-GMM изчислена чрез правилото на Bayes. С *r* е отбелязан relevance factor (коефициент на влияние) – когато този коефициент нараства адаптираните средни вектори остават близки по стойност до средните вектори на UBM μ_i^{UBM} .

1.2.2.2.Метод на опорните вектори

Метода на опорните вектори [Zou et al., 2014] е двоичен линеен класификатор, където оптималната разделяща хипер равнина се дефинира от опорните вектори. Ако класовете не са линейно разделими, то се избира съответната функция на ядрото, която да трансформира входните вектори с признаци в ново признаково пространство, в което преобразуваните вектори да бъдат вече линейно разделими.

Ако тестовият вектор е y то решаващата функция на SVM е дефинирана като

$$f(y) = \operatorname{sgn}(\sum_{i=1}^{L} a_i t_i K(x_i, y) + b) \tag{1.15}$$

където t_i е идеалният изход, *L* е броят на опорните вектори, x_i е *i*-тия опорен вектор, a_i е тегловният коефициент на *i*-тия опорен вектор, *b* е константно изместване и *K(.)* е функция на ядрото. Идеалния изход при VAD–алгоритмите е 1 за векторите с говор и -1 за тези с шум. Като функция на ядрото често се използват линейна функция и радиална базисна функция (radial basis function - RBF) във вида

$$K_{lin}(x, y) = < x, y > \tag{1.16}$$

$$K_{RBF}(x, y) = \exp(-\gamma \left\| x - y \right\|^2) \tag{1.17}$$

където <.> е скаларно произведение, γ е параметър (ширина на ядрото), а $\|.\|$ е Евклидово разстояние.

В [Ramirez et al., 2006] е показано, че в сравнение с линейната функция използването на RBF при VAD-алгоритмите води до по-добри резултати. При тестване

входния вектор *y* ще бъде класифициран като говор ако решаващата функция *f(y)≥1* и като шум в обратния случай.

1.2.2.3. Метод с i-вектори

I-векторите са вектори с ниска размерност базиращи се на средните GMM супер-вектори, адаптирани чрез UBM и получени от тях чрез факторен анализ. При този подход се предполага, че специфичната за даден диктор информация е разположена в под-пространство с ниска размерност [Hasan et al., 2014].

В [Dehak et al., 2011] е предложен подход, в който е дефинирано едно общо подпространство наречено '*пространство на общата изменчивост'* (total variability space) съдържащо едновременно изменчивостта на диктора и канала. То се дефинира чрез total variability matrix (матрица на общата изменчивост) която съдържа собствени вектори с най големи собствени стойности на total variability covariance matrix. За дадено произнасяне s зависещия от диктора и от канала GMM супер-вектор се дефинира като

$$m_s = m_0 + Tw_s \tag{1.18}$$

където m_0 е супер-вектор независим от диктора и от канала (получен чрез UBM), T е правоъгълна матрица от нисък ранг и w_s е случаен вектор с нормално разпределение $N(0, I)$ чийто компоненти са известни като пълни фактори (total factors) [Dehak et al., 2011]. Апостериорния среден вектор на w_s за дадено произнасяне се разглежда като i-вектор (identity vector) [Hasan et al., 2014].

Вектора w е скрита променлива, която може да бъде дефинирана чрез нейното апостериорно разпределение с използване на статистиката на Baum-Welsh [Dehak et al., 2011]. Ако y е последователност с дължина L сегмента $\{y_1, y_2, ..., y_L\}$ и UBM Ω съдържа C компонента дефинирани в признаково пространство с размерност F , то Baum-Welsh статистиките имат вида

$$\begin{aligned} N_c &= \sum_{t=1}^{L} P(c \mid y_t, \Omega) \\ F_c &= \sum_{t=1}^{L} P(c \mid y_t, \Omega) y_t \end{aligned} \tag{1.19}$$

където $c = 1, ..., C$ са индексите на Гаусовите компоненти и $P(c \mid y_t, \Omega)$ съответства на апостериорната вероятност на компонента c генерираща вектора y_t . За да се определи i-векторът е необходимо да се изчислят центрираните статистики на Baum-Welsh от първи ред на базата на средната стойност на компонентите на UBM

$$\tilde{F}_c = \sum_{t=1}^{L} P(c \mid y_t, \Omega)(y_t - m_c) \tag{1.20}$$

където m_c е средната стойност на компонентата *c* или i-векторът за дадено произнасяне се определя като

$$w = (I + T^t \sum^{-1} N(u)T)^{-1} T^t \sum^{-1} \tilde{F}(u) \tag{1.21}$$

N(u) е диагонална матрица с размерност *CF x CF*, чийто диагонални блокове са $N_c I$. $\tilde{F}(u)$ е супер вектор с размерност *CFx1* получен чрез конкатенация на всички Baum-Welsh статистики от първи ред $\tilde{F}_c$ за даденото произнасяне *u*. $\sum$ е диагонална ковариационна матрица с размерност *CF x CF* определена по време на факторния анализ и тя моделира остатъчната изменчивост необхвaната от матрицата на тоталната изменчивост *T* [Dehak et al., 2011].

При сравнение на два i-вектора често се използва метода на опорните вектори с косинусно ядро. При косинусното ядро се взема предвид само ъгъла между двата вектора, но не и техните амплитуди. Предполага се, че информацията за канала и сесиите влия върху амплитудата на i-векторите и елиминирайки тази информация се увеличава робастността на цялата система [Dehak et al., 2009].

След като са определени i-векторите, за да се компенсира влиянието на канала те се подлагат на нормализация по дължина [Garcia-Romero et al., 2011]. С тези вектори е тестван в [Khoury et al., 2016] PLDA (вероятностен линеен дискриминантен анализ - Probabilistic Linear Discriminant Analysis – PLDA) подход при VAD-алгоритъм. Изчисляването на LLR h_{PLDA} за тестовия клъстер C_t принадлежащ към клас говор е във вида

$$h_{PLDA}(C_t) = \frac{p(w_t, w_{speech} \mid \Theta)}{p(w_t \mid \Theta)\, p(w_{speech} \mid \Theta)} \tag{1.22}$$

където w_t е тестовият i-вектор, w_{speech} е средният i-вектор за клас ‘говор’ и $\Theta = \{F, G, \sum_{\varepsilon}\}$ е PLDA модела. *F* и *G* са между-класовата и вътрешно-класовата ковариационни матрици и $\sum_{\varepsilon}$ е ковариационната матрица на остатъчния шум (residual noise)

1.2.3. VAD-алгоритми в системи за разпознаване на диктори

В тази подточка ще бъдат разгледани значими (според автора) разработки на VAD-алгоритми, използвани в системи за разпознаване на диктори. Описани са подробно използваните класификационни схеми и бази данни с говор, получените

експериментални резултати и направената оценка на грешките (при детекция и при разпознаване).

1.2.3.1. Изследване на VAD-алгоритми при независима от текста верификация на диктори в NIST SRE

В работата [Mak et al., 2014] са предложени VAD-алгоритми специално адаптирани за NIST 2010 SRE. Особеността при тези тестове за верификация на диктори е качеството на говорния сигнал. Част от записите са интервюта, записани с различни микрофони и от разстояние (far-field microphones). В една немалка част от тях отношението сигнал/шум (Signal-to-Noise Ratio - SNR) е около 5 dB. Налице е също така и значително количество пикове в сигнала. Наблюдава се периодичен фонов шум маскиращ нискоенергийните фрагменти на говорния сигнал, а в някои записи има и смесване на разговорите на интервюиращия и интервюирания (cross-talk).

Използвани са два алгоритъма за VAD – алгоритъм на Sohn [Sohn et al., 1999] базиран на статистически модел и алгоритъм на Fukuda [Fukuda et al., 2010] прилагащ модел с Гаусови смеси.

1.2.3.1.1. VAD алгоритъм на Sohn

Детекцията на говор се разглежда като класическа задача за проверка на хипотези. Дефинират се хипотезите H_0 и H_1

$$\begin{aligned} H_1 &: speech\ present : Y(m) = X(m) + B(m) \\ H_0 &: speech\ absent : Y(m) = B(m) \end{aligned} \tag{1.23}$$

където $Y(m), X(m)$ и $B(m)$ са съответно DFT преобразуванията на зашумения говор, чистия говор и фоновия шум за $m^{-я}$ сегмент. Предполага се, че комплексните коефициенти в тях са независими и нормално разпределени. За всеки сегмент се изчислява статистическата оценка $\Gamma(m)$

$$\Gamma(m) = \frac{P(H_0)}{P(H_1)}\left[\frac{a_{01} + a_{11}\Gamma(m-1)}{a_{00} + a_{10}\Gamma(m-1)}\right] \begin{matrix} \Lambda(m) > \eta : H_1 \\ \leq \eta : H_0 \end{matrix} \tag{1.24}$$

където: $a_{ij} \triangleq \Pr(q(m)) = H_j \mid q(m-1) = H_i)$ са преходни вероятности; $P(H_0)$ и $P(H_1)$ са начални вероятности и $\Lambda(m)$ е отношението на правдоподобие (likelihood ratio - LLR). Тъй като се предполага, че DFT коефициентите са независими то

$$\Lambda(m) = \left[\prod_{k=0}^{K-1} \frac{p(Y_k(m) \mid H_1)}{p(Y_k(m) \mid H_0)}\right]^{1/K} \tag{1.25}$$

където K е броят на дискретните честоти и p() са функции на плътността на вероятността на нормално разпределени комплексни величини (complex normal densities).

Детекцията на говорните фрагменти се реализира като статистическата оценка $\Gamma(m)$ се сравнява с фиксиран праг η. За да се определи този праг се сортират в низходящ ред стойностите на $\Gamma(m)$ в анализирания файл. Определя се средната стойност $\bar{\Gamma}_b$ на крайните 10% в така получената последователност (считат се за фонов шум). Освен това се намира минималната стойност в началните 5% от последователността (считат се за пикове в сигнала). Праговата стойност η се изчислява като

$$\eta = v\bar{\Gamma}_b + (1-v)\min[\Gamma(p_1),\ldots,\Gamma(p_L)] \tag{1.26}$$

където v е тегловен коефициент.

1.2.3.1.2. VAD алгоритъм на Fukuda

Описания в [Fukuda et al., 2010] алгоритъм прилага при детекция на говор модела на Гаусови смеси (Gaussian mixture model - GMM). При този модел за $m^{-я}$ сегмент се изчислява LLR $L(m)$ и решението говор/не-говор се приема съгласно

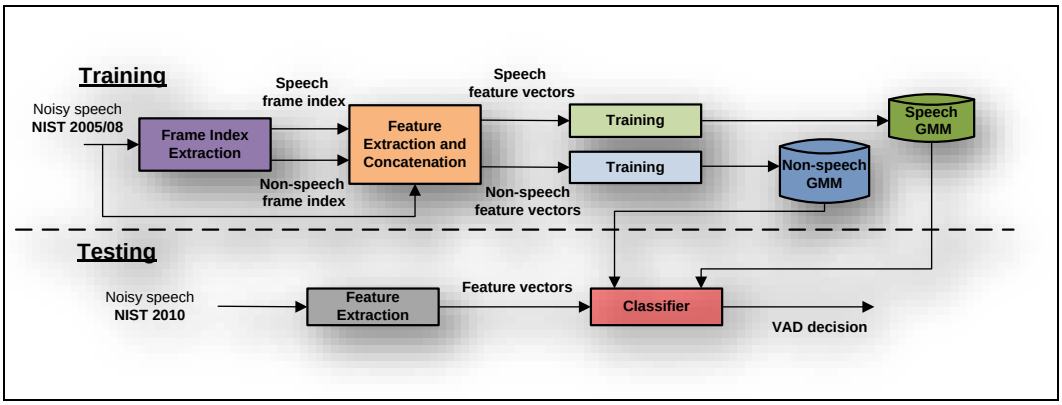
$$\begin{aligned} L(m) = \log(p(y(m)\mid H_1)) - \log(p(y(m)\mid H_0)) &> \eta : H_1 \\ &\le \eta : H_0 \end{aligned} \tag{1.27}$$

Предполага се, че параметричните вектори y могат да бъдат представени чрез смес от Гаусови разпределения

$$p(y\mid H_i) = \sum_{j=1}^{J} w_{ij}\mathrm{N}(y;\mu_{ij},\Sigma_{ij}) \tag{1.28}$$

$w_{ij}, \mu_{ij}, \Sigma_{ij}$ са тегловни коефициенти, средни вектори и ковариационни матрици на Гаусовите модели съответно при говор ($i=1$) и при не-говор ($i=0$).

Като признаци във VAD алгоритъма са използвани Мел кепстралните коефициенти, но допълнително обработени така, че в тях да се подсили влиянието на хармоничната структура на спектъра (harmonic structure-based mel cepstral coefficients) [Fukuda et al., 2010].



Фиг. 1.2. Блок схема на етапите на обучение и тестване при GMM VAD-алгоритъм (Фигурата е адаптирана версия на фиг. 8 в [Mak et al., 2014]).

В алгоритъма са включени два GMM (говор и не-говор) за чиито обучение са необходими предварително подготвени данни. В конкретния случай това трябва да бъдат сегментирани в два класа говорни последователности, които са част от NIST SRE. Файлове с подобна сегментация обаче не са включени в NIST SRE, което от своя страна е наложило авторите да разработят алгоритъм за груба предварителна сегментация (Frame Index Extraction). За да се намали нивото на шума в сигнала се прилага допълнително спектрално изваждане. На фиг. 1.2 е показана блок схема на етапите на обучение и тестване на GMM VAD-алгоритъма [Mak et al., 2014].

1.2.3.1.3. Оценка на ефективността на VAD алгоритмите

При оценка на ефективността на VAD алгоритмите най-често се извършва сравнение - чрез Receiver Operating Characteristics (ROC) [Fawcett, 2006] графики - между резултатите от ръчната сегментация (реализирана върху незашумената версия на сигнала) и тези получени от прилагането на разработвания алгоритъм върху зашумената версия на сигнала. Тъй като в NIST SRE не са включени незашумени версии на аудио записите, то оценката на ефективността на разработените VAD алгоритми се извършва косвено – чрез оценка на влиянието на тези алгоритми върху точността на верификация на диктори. Определят се различните грешки при верификация и техните графични интерпретации, а именно стойността на равно вероятната грешка - Equal Error Rate (EER), Detection Error Tradeoff (DET) графики и минималната стойност на функцията на грешката от детекция – minimum Decision Cost Function (minDCF) [Beigi, 2011].

1.2.3.1.4. Базова система за верификация на диктори

При верификация на диктори са тествани две системи. При първата е приложен GMM-SVM подхода [Campbell et al., 2006a], а при втората метода с i-вектори [Dehak et al., 2011]. Като признаци при GMM-SVM подхода са използвани 12 Мел кепстрални коефициента и техните първи производни. Допълнително е приложено върху тях CMS (Cepstral Mean

Subtraction – центриране на последователността от кепстрални вектори) и feature warping [Beigi, 2011]. Създаден е универсален фонов модел (Universal Background Model - UBM) с 512 компонента (mixtures) чрез използване на записите с интервюта от NIST 2005-06. При обучение на GMM за target (целеви) диктори е приложена адаптация на Bayes с relevance factor 16. Същата адаптация е използвана за обучение на моделите на 300 impostors (нецелеви) диктори (background dataset). Техните средни вектори се конкатенират, за да се получи GMM-супер вектор с размер 12288. Модела GMM-SVM за всеки target диктор е обучен чрез target-dependent GMM супер-вектор и background GMM супер-вектори. За да се намали влиянието на комуникационния канал е използван подхода с nuisance attribute projection (NAP) матрица, която е изчислена върху говорен материал от 144 диктори от NIST 2005-08 [Campbell et al., 2006b].

При система с i-вектори като признаци са използвани 19 Мел кепстрални коефициента и техните първи и втори производни. Допълнително върху тях е приложено CMS и feature warping [Beigi, 2011]. Данни от NIST 2006-08 са използвани за обучение на UBM с 1024 компонента. Определена е матрицата на пълната изменчивост (total variability matrix) - с 400 базисни колони (factors) на базата на 6102 произнасяния от 191 диктора от NIST 2005-08. Данни от тези диктори са използвани за определяне на матрицата на факторните тегла (loading matrix) със 150 латентни променливи при Гаусов вероятностен линеен дискриминантен анализ (Gaussian Probabilistic Linear Discriminant Analysis – G-PLDA) [Garcia-Romero et al., 2011]. Приложена е нормализация на дължината на всички i-вектори преди изчисляване на матрицата на факторните тегла [Garcia-Romero et al., 2011].

1.2.3.1.5. Експериментални резултати

В работата са тествани следните VAD-алгоритми. Това са: AE-VAD (използва се енергията на сигнала), ASR-VAD (сегментирани данни, получени от система за разпознаване на говор и предоставени от NIST [NIST SRE, online]), GMM-VAD (алгоритъм използващ модел с Гаусови смеси [Fukuda et al., 2010]), SM-VAD (алгоритъм на Sohn [Sohn et al., 1999]), SS+SM-VAD (SM-VAD с използване на спектрално изваждане), SS+AE-VAD (AE-VAD с използване на спектрално изваждане).

При верификация на диктори със системата GMM-SVM детектора SM-VAD се представя по-добре от GMM-VAD за записи с интервюта в NIST SRE. Основната причина е необходимостта от голямо количество предварително сегментиран говор за обучението на GMM. Прилагането на спектралното изваждане силно подобрява точността при детектора използващ енергията на сигнала – AE-VAD и слабо влияе върху

точността на SM-VAD. При статистическия модел нивото на фоновия шум е взето предвид при изчисляване на оценъчната функция и в този случай спектралното изваждане не е достатъчно ефективно. Най-добри резултати и при двата критерия – EER и minDCF - са получени при SS+AE-VAD.

При верификация на диктори със системата използваща i-вектори са тествани четири версии на SM-VAD. Предполага се, че разпределението на коефициентите на Фурие може да бъде съответно: на Гаус (базов алгоритъм), на Лаплас и Гама разпределение. Четвъртият тест е с Гаусово разпределение, но е използвано спектрално изваждане на етапа на предварителната обработка. Експериментите показват, че SM-VAD с Гама разпределение демонстрира по-добри резултати от базовия алгоритъм при критерий EER, но ако критерият е minDCF то резултатите са обратни. Спектралното изваждане има ефект при базовия алгоритъм, но само при критерий EER.

1.2.3.2. Изследване на VAD-алгоритми при зависимо от текста разпознаване на диктори с корпуса RSR2015

В работата [Alam et al., 2014] са описани алгоритми за детекция на говор предназначени за зависима от текста верификация на диктори с корпуса RSR2015 [Larcher et al., 2012]. Тези VAD-алгоритми използват съответно: енергията на говорния сигнал [ISIP, online], енергията на сигнала, но с допълнителна корекция (hangover scheme), векторно квантуване [Kinnunen et al., 2013], модел с Гаусови смеси с последователна адаптация (sequential Gaussian mixture model) [Ying et al., 2011], supervised GMM и unsupervised GMM.

1.2.3.2.1. VAD-алгоритъм използващ енергията на сигнала

В работата [ISIP, online] първо се изчислява логаритмичната енергия на сигнала за всеки сегмент. Получения енергиен контур се изглажда чрез филтър с изместваща се средна стойност. Стойностите на контура се сортират по големина и тези от тях, които съответстват на 20% и на 80% от дължината на сортираната последователност се приемат за първа и втора прагови стойности. Тяхната средна стойност дефинира енергиен праг, чрез който се сепарират сегментите говор/шум. Така полученото предварително решение за вида на даден сегмент допълнително се коригира (изглажда) чрез hangover алгоритъм. Този алгоритъм коригира по два различни начина приетото вече решение. При прехода от шум към говор се забавя (отлага) решението за конкретен сегмент докато не се установи доминиращата стойност в анализирания hangover буфер. Ако стойността в буфера е говор - то и конкретния сегмент ще бъда говор. По този начин се избягва грешка от вида “шум приет като говор“ (false alarm error - FAE). При прехода от говор към шум

се въвежда подобно забавяне на решението с цел да се намалят грешките от вида “говор приет като шум“ (miss detection error - MDE). В конкретния алгоритъм допълнително са определени граничните точки на анализираното изречение чрез анализ на продължителността на паузите.

1.2.3.2.2. VAD-алгоритъм чрез векторно квантуване

Предложения VAD-алгоритъм е със самообучение (unsupervised) и за разлика от GMM алгоритъма, разгледан в предходния параграф не изисква предварително сегментирани и етикирани говорни данни [Kinnunen et al., 2013].

Алгоритъма включва следните етапи. С цел подобряване на SNR на говорния сигнал първо се прилага спектрално изваждане, след което се изчислява логаритмичната енергия за всеки сегмент. Стойностите й се сортират във възходящ ред и 10% от индексите на сегментите в двата края оформят съответно групите на сегменти с шум и с говор. Мел кепстъра се изчислява за всеки сегмент, но от необработения говорен сигнал (без спектрално изваждане). Чрез K-means алгоритъма се формират кодовите книги (codebooks) на говора и шума, като се използват кепстралните коефициенти на сегментите от вече определените две групи. При детекцията на текущия сегмент се използва Евклидово разстояние. Крайното VAD-решение се получава чрез прилагане на hangover scheme и алгоритъм за определяне на граничните точки.

1.2.3.2.3. *VAD-алгоритъм използващ модел с Гаусови смеси с последователна адаптация*

Предложения алгоритъм е със самообучение [Ying et al., 2011]. Спектъра на говорния сигнал се разделя на 8 Мел честотни ленти. Изчислява се логаритмичния спектър в тях и се извършва филтрация във времевата област чрез медианен филтър.

Използвания Гаусов модел съдържа две разпределения. Първоначално модела се обучава с първите няколко десетки сегменти от произнасянето. Тези сегменти се клъстеризират в две разпределения като това с по-малка средна стойност съответства на фрагментите с шум, а с по-голяма – на говорните. Праговата стойност за детекция на говор/шум се определя в точката между двете разпределения, където вероятностите са еднакви. Детекцията на даден сегмент се реализира поотделно за всяка честотна лента като решението за сегмента се приема чрез гласуване (voting procedure). Крайното VAD-решение, аналогично на описания по-горе алгоритъм, се получава чрез прилагане на hangover scheme и алгоритъм за определяне на граничните точки.

1.*2.3.2.4. VAD-алгоритъм използващ модел с Гаусови смеси и обучение с учител*

В работата [Alam et al., 2014] моделите с Гаусови смеси се обучават със записан по телефона говор от NIST SRE 2004-2010. Определя се Мел кепстъра с 11 коефициенти,

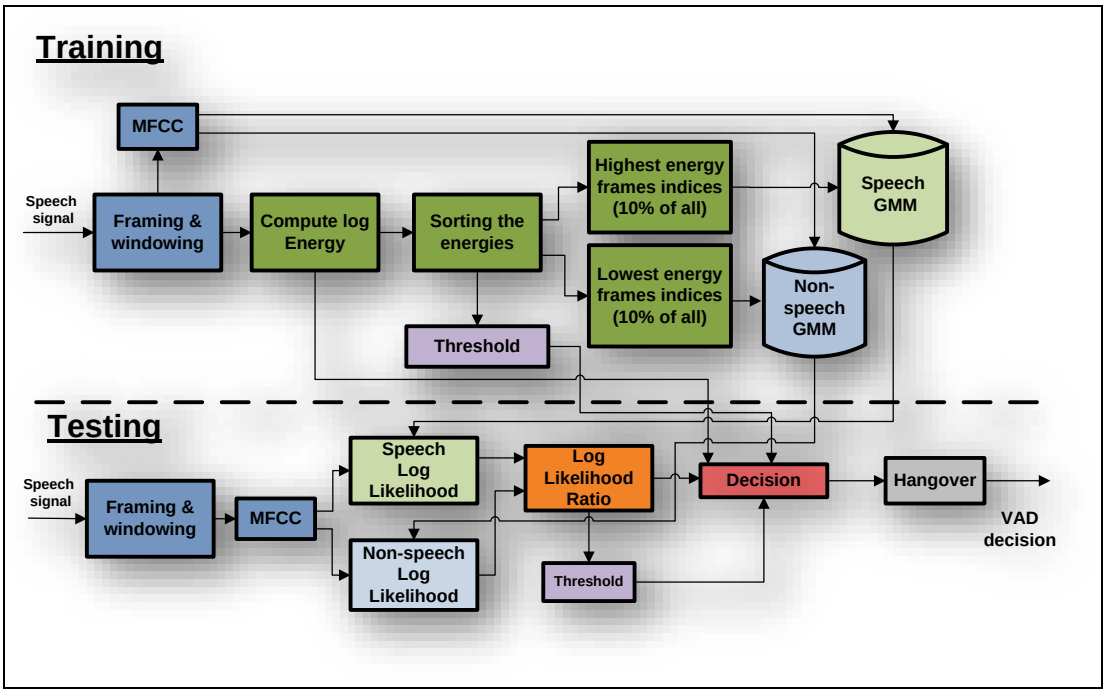
както и първата, втората и третата му производни. За обучение се използват предварително сегментирани и етикирани говорни данни. Обучени са 2 модела с Гаусови смеси с по 256 компонента (mixtures) и диагонални ковариационни матрици съответно за данните с говор и с шум. Детекцията на говорните сегменти се реализира с данни от корпуса RSR2015 в следната последователност: определяне на Мел кепстрален вектор с 44 коефициента; изчисляване на LL (log likelihood) за всеки вектор спрямо всеки един от двата модела; прилагане на медианна филтрация върху двата LL контура; изчисляване LLR и сравняване с прагова стойност.

1.2.3.2.5. VAD-алгоритъм използващ модел с Гаусови смеси и самообучение

Разгледания в този параграф алгоритъм [Alam et al., 2014] е принципно подобен на описания в т. *2.3.2.2* алгоритъм използващ ВК. При ВК моделите говор/шум се получават чрез прилагане на k-means клъстеризация. При модела с Гаусови смеси и самообучение, k-means се използва само при инициализация. Детекцията на сегментите говор/шум при разглеждания алгоритъм се реализира в следните основни етапи:

- определяне на логаритмичната енергия на сигнала за всеки сегмент, сортиране на стойностите на енергията, формиране на групите сегменти за шум и за говор, определяне на енергиен праг аналогично на този в т.2.3.2.1.
- обучение на GMM за всеки клас с Мел-кепстъра от съответната група сегменти, определяне на LLR за всеки сегмент, изглаждане на LLR контура, определяне на LLR прага; детекция на говорните сегменти чрез използване на два прага – енергиен и LLR праг;
- прилагане на hangover scheme;
- прилагане на алгоритъм за определяне на граничните точки за получаване на крайното VAD-решение.

На фиг. 1.3 е показана блок схема на VAD-алгоритъма използващ GMM със самообучение и описан в [Alam et al., 2014].



Фиг. 1.3. Блок схема на VAD-алгоритъм използващ GMM със самообучение.
(Фигурата е адаптирана версия на фиг. 3 в [Alam et al., 2014]).

1.2.3.2.6. Корпус RSR2015

Корпуса с говорни данни RSR2015 е предназначен за зависимо от текста разпознаване на диктори. Той съдържа аудио записи от 298 диктора (142 жени и 156 мъже), всеки от тях записан в 9 сесии или общо корпуса съдържа говор с продължителност от 151 часа.

Записите за направени в офис чрез използване на 6 мобилни устройства (смартфони и таблети). Данните са организирани в три групи. Първа група съдържа 30 изречения от базата данни TIMIT с единична дължина 3.2 секунди и обща продължителност на записите - 71 часа. Втората група включва 30 кратки команди с единична дължина 2 секунди и обща продължителност - 45 часа. Третата група съдържа произнасяния на последователности от цифри съответно 3 последователности в 10 сесии и 10 последователности в 5 сесии. При тестовете в работата са използвани записи само от първата група.

1.2.3.2.7. Базова система за зависима от текста верификация на диктори

Система за верификация на диктори се базира на GMM UBM с 512 компонента и диагонална ковариационна матрица. Говорните данни, използвани за обучение на UBM включват 63587 записа от 97 диктора (мъже и жени). От Мел кепстралните коефициенти и обучения UBM се определя статистиките на Baum-Welch. Тези статистики се използват при метода на съвместния факторен анализ (Joint Factor Analysis - JFA). В работата е използвана системата за верификация описана в [Kenny et al., 2014].

1.2.3.2.8. Експериментални резултати

Тествани са шест VAD-алгоритъма като за всеки от тях е реализирана отделна верификация на диктори. Алгоритмите са описани по-горе в текста като тези използващи енергийния контур са два – оригиналния описан в [ISIP, online] и модификацията предложена от авторите. Получените резултати показват, че двете версии на GMM VAD-алгоритъма (с учител и със самообучение) водят до по-ниска грешка при верификация в сравнение с всички останали. Алгоритъма със самообучение се представя най-добре от гледна точка на използваните критерии за оценка на грешката– EER и minDCF.

1.2.3.3. VAD-алгоритми на базата на невронни мрежи

В този параграф ще бъдат разгледани VAD-алгоритми използващи многослоен перцептрон (multi-layer perceptron - MLP) [Ganapathy et al., 2011] и невронна мрежа с дълбочинно обучение [Dwijayanti et al., 2018].

1.2.3.3.1. VAD-алгоритъм на базата на многослоен перцептрон

Предложения алгоритъм се базира на апостериорните вероятности на фонемите в английския език, получени на изходите на невронна мрежа. При обучение на MLP са използвани признаци, получени чрез метода на линейно предсказване в честотната област (frequency domain linear prediction - FDLP) [Ganapathy et al., 2010]. Чрез него се формира 420 размерен вектор на признаците.

Обучението на MLP е реализирано с данни от корпуса CTS (conversational telephone speech) [Hain et al., 2005] който съдържа телефонни разговори с продължителност 180 часа. От тях 100 часа са използвани за обучение и 30 часа за тест. Фонетичната транскрипция на данните (44 фонемни класа плюс клас шум) е получена предварително чрез HMM-GMM система, описана в [Hain et al., 2005]. Използвания многослоен перцептрон е с един скрит слой с 5000 неврона и изходен слой с 45 неврона (softmax). Изходния вектор с размерност 45 се преобразува във вектор с размерност 2 чрез сумиране на вероятностите за всички фонеми, т.е. получава се класификатор с два изхода (говор и шум). Решението говор/шум се получава чрез прилагане на фиксиран праг.

Верификацията на диктори е на базата на системата GMM-UBM с i-вектори и GPLDA [Garcia-Romero et al., 2011]. За да се обучи UBM и да се получи матрицата на пълната изменчивост при определяне на i-векторите са използвани данни от NIST 2004 SRE, Switchboard II Phase III и NIST 2006 SRE. При обучение е използван VAD-алгоритъм, предоставен от NIST. При верификация на диктори тестове са реализирани със следните VAD-алгоритми – с адаптивна енергия на сигнала [Reynolds et al., 2005], с

Мел-кепстър, с времево-честотна модулация [Mesgarani et al., 2006], MLP1 - предложения алгоритъм, но като признаци е използван Мел-кепстъра с CMS и MLP2 - предложения алгоритъм, но признаците са получени с FDLP. Точността на верификация се оценява чрез EER, а точността на детекция чрез общата средна стойност на грешките FAE и MDE изчислена за всички произнасяния. Експерименталните резултати показват, че най-висока точност при верификация на диктори е постигната при използване на VAD-алгоритъма MLP2. Интересен е фактът, че при MLP2 минимална EER се получава дори когато обучението и теста са реализирани на различни езици. Освен това при сравнение на точността на детекторите отново минимална грешка се получава при MLP2. Детайлния анализ на резултатите показва, че в сравнение с другите детектори при този детектор е налице значително по-малък брой грешки от вида false alarm.

1.2.3.3.2. VAD-алгоритъм използващ невронна мрежа с дълбочинно обучение и производни на спектъра на мощността

В работата [Dwijayanti et al., 2018] е предложен VAD-алгоритъм на базата на DNN, при който се използва предварително обучение (без учител) на невронната мрежа. Архитектурата на предложената DNN се състои от 5 слоя като всеки от тях съдържа невронна мрежа от вида ‘ограничена машина на Bolzmann‘ (restricted Bolzmann machine - RBM) с бинарни неврони във видимия и скрития слой. Обучението на RBMs се извършва последователно (stacked RBMs). След като дадена RBM е обучена изходните стойности на невроните в скрития й слой се използват като признаци за обучение на следващата RBM. При обучението на RBM е приложен алгоритъма за градиентно спускане с контрастивна дивергенция. След приключване на обучението на всички RBMs се извършва фина настройка на теглата на невроните на DNN, като се използва алгоритъма за обратно разпространение на грешката приложен върху цялата мрежа.

В предварителни изследвания са сравнени резултатите от детекция получени при използване на т.н. първични (raw) и вторични (hand-crafted) признаци. Като първични признаци са избрани логаритмичния спектър на мощността и неговите първа и втора производни по честота (делта-спектър), а като вторични Мел-кепстъра и съответно делта и делта-делта-кепстъра. Оказва се, че при първичните признаци резултатите от детекция са по-добри и освен това използването на делта параметри води до допълнително подобрение.

При VAD-алгоритъма признаците на входа на DNN се получават по следния начин. Първо, с цел намаляване нивото на шума говорния сигнал се обработва чрез метода с изместващи се спектрални филтри (running spectral filters – RSF [Fujioka et al.,

2006]). Второ, чрез евристични правила се формират бинарни последователности (masks) на базата на логаритмичния спектър на мощността получен на изхода на RSF и на неговите първа и втора производни. Трето, бинарните последователности се умножават със спектъра и се получават т.н. вероятни говорни фрагменти (speech period candidates - SPC). SPCs заедно с логаритмичния спектър на мощността се подават на входа на DNN.

Тестове са реализирани с говор от корпуса ASJ Continuous Speech Corpus for Research vol. 2 [SRC, online]. Зашумените версии на записите са получени чрез добавяне към чистия сигнал на шум от корпуса NOISEX-92 [NOISEX, online]. Броят на невроните в RBMs в 5-те слоя на DNN са съответно – 200, 200, 200, 200 и 100. Количествената оценка на резултатите се извършва чрез ROC анализ. Като критерий е избрана площта под съответната ROC графика (area under the curves – AUC [Fawcett, 2006]). Експериментите показват, че чрез комбинация на параметъра SPC и логаритмичния спектър се получават по-добри резултати от тези при самостоятелното използване на спектъра. Авторите демонстрират чрез експерименти, че в сравнение с алгоритмите на Sohn [Sohn et al., 1999], [Ramirez et al., 2004] и Kinnunen [Kinnunen et al., 2013] при предложения от тях VAD-алгоритъм са налице по-добри резултати особено при нестационарен шум и ниско SNR. Независимо че описания в работата детектор не е част от система за разпознаване на диктори той е включен в обзора по една единствена причина. А тя е, че при него се демонстрира възможността на DNN чрез статистическо обучение със значителни по обем данни да се постигне ефективно разделяне на класовете чрез използване само на първични (raw) параметри.

1.3. Определяне на гранични точки на кратки фрази

1.3.1. Увод

Определянето на граничните точки на говорните фрагменти е етап от предварителната обработка при автоматичните системи за разпознаване (на говор и на диктори) и резултатите получени на този етап влияят съществено върху крайната точност на разпознаване. Необходимо е да се посочи, че редица алгоритми, използвани при робастно разпознаване на диктори като например CMS изискват точно определяне на граничните точки на говорното съобщение с цел изчисляване на средния кепстрален вектор само чрез кепстъра на думата или фразата.

Известно е, че определянето на граничните точки се затруднява съществено при предаване на говор по комуникационен канал. С навлизането на безжичните комуникации и VoIP (Voice over Internet Protocol) този проблем придоби изключителна

актуалност. Главната причина е, че при подобен вид предаване на данни е налице повече шум (от кодиране, от загуба на пакети от импулси и др.), отколкото при стационарните телефони [Muralishankar et al., 2010].

Особеностите на даден алгоритъм за определяне на граничните точки зависи от областта, в която той се използва. Всяка една област (например - разпознаване на говор, разпознаване на диктори, кодиране на говор и др.) налага различни изисквания за сложност, робастност, точност, време за реакция и др. Независимо от факта, че в областта на ОГТ се работи от няколко десетилетия все още не е създаден универсален алгоритъм.

1.3.2. Алгоритми за определяне на гранични точки

Известни са три подхода при определяне на граничните точки – експлицитен, имплицитен и хибриден [Roy, 2019]. При експлицитния подход определянето на граничните точки предхожда и е независимо от класификационната схема, използвана в разпознаващата система. При имплицитния подход граничите точки се определят в процеса на класификация, т.е. няма отделен етап за тяхното определяне. Хибридния подход съчетава идеи от предходните два.

По-долу в текста под алгоритми за ОГТ се разбират алгоритми основаващи се единствено и само на експлицитния подход, освен ако изрично не е посочено друго.

Един алгоритъм за ОГТ включва два основни етапа - извличане на признаци и приемане на решение. На първия етап се изчисляват един или няколко признака на говорния сигнал, например: енергия на сигнала [Li et al., 2012], спектрална ентропия [Wu et al., 2005], [Zhang et al., 2013], [Zhang et al., 2016], времево-честотни параметри [Kyriakides et al., 2011], s-трансформация [Xunbo et al., 2018], МЕЛ-кепстър [Cao et al., 2017], уейвлети [Yali et al., 2014], [Zhang et al., 2013] и др. На втория етап, най-често се използва или краен автомат [Li et al., 2002], [Chung et al., 2014] или класификатор - с невронни мрежи [Comas et al., 2005], [Wu et al., 2012], скрити Марковски модели - HMM [Zhang et al., 2005], метод на опорните вектори - SVM [Feng et al., 2016] и др.

Прилагането на подходи от областта на разпознаване на образи и машинно обучение при ОГТ има както предимства, така и ограничения. Главното предимство в случая е получаване на висока точност, но постигната с цената на въвеждане на етап на обучение. За целта е необходимо да се подготвят говорни данни за отделните класове, което от своя страна най-често изисква ръчна сегментация на говорния материал. Посоченото ограничение затруднява значително автоматизирането на целия процес на определяне на граничните точки [Wu et al., 2012].

Както бе посочено структурата на даден алгоритъм за ОГТ зависи както от вида на решаваната задача, така и от особеностите на разпознаващата система, в която той участва. Поради голямото разнообразие на признаци и класификатори използвани в алгоритмите за ОГТ, в представения обзор е решено да бъдат описани само най-характерните от тях, намиращи приложение в областите зависимо от текста разпознаване на диктори и разпознаване на думи и фрази предимно в зашумена среда.

1.3.2.1. Изследване на Comas с четири класификатора

Реализирано е експериментално изследване за ОГТ с четири различни признака и четири различни класификатора [Comas, 2005]. Признаците за всеки сегмент са: енергията на Teager, нейната производна, спектрална ентропия и спектрална кохеретност между съседни сегменти. Класификаторите са: линеен дискриминантен, AdaBoost при линейни класификатори, bagging при класификатори с дървовидна структура и многослоен перцептрон (МСП). При разпознаване на отделния сегмент и при тестове със зашумени данни при всички класификатори е налице високо ниво на False Positive Rate (отношение на грешно класифицираните не-говорни сегменти към всички не-говорни сегменти). Това означава, че при сегментите има силно застъпване между класовете говор/не-говор. Използването на bagging при класификатора с дървовидна структура подобрява точността. Краен автомат с четири състояния формира окончателното решение за даден сегмент чрез анализ на предходните решения на класификаторите. Максимална точност при ОГТ на зашумен сигнал е постигната при МСП – 84%. Повечето грешки според авторите се получават в преходните области между двата класа (говор и не-говор).

1.3.2.2. Алгоритъм на Wu използващ адаптивна лентова спектрална ентропия

В работата [Wu et al., 2005] се предлага спектъра на говорния сигнал да бъде разделен на 32 еднакви честотни ленти. Целта на подобно разделяне е да се елиминират честотните ленти, които най-често съдържат шумови компоненти. Тази идея се прилага за спектралната ентропия и се формира нов параметър – Band-partitioning Spectral Entropy (BSE). Допълнително е предложен алгоритъм за адаптивно шумопотискане, който позволява да бъдат избирани адаптивно във времето подходящите честотни ленти. Този алгоритъм е наречен – Refined Adaptive Band Selection (RABS). В работата е формиран нов робастен параметър, който комбинира BSE параметъра и RABS алгоритъма, а именно – Adaptive Band-partitioning Spectral Entropy (ABSE). Той е тестван при детекция на говор в реална зашумена среда (adverse environments). Въведен е адаптивен праг за детекция на говорен/шумов сегмент на базата на статистическите параметри на логаритмичната стойност на ABSE. Тестове са реализирани както за говор с добавен

адитивен шум, така и за говорни реализации, записани в движеща се кола при наличие на музикален фон. Получените резултати показват, че предложените параметър и алгоритъм за ОГТ се справят успешно както при адитивен, така и при реален фонов шум и постигат вероятност 89.2% за правилна и 3.5% за грешна детекция на говорните сегменти.

1.3.2.3. Алгоритъм на Kyriakides използващ анализ на спектрограми

В работата [Kyriakides, 2011] е предложен нов подход за ОГТ на изолирани думи наречен Variance Kernel Method (VKM). При него се търсят области от спектрограмата на говорния сигнал, които се отличават със значителна локална дисперсия. Счита се, че тези области съответстват на говорни фрагменти. Спектрограмата се обработва като изображение с максимално застъпващи се филтри 5x5. В областите локализирани от тези филтри се определя средно квадратичното отклонение в стойностите на пикселите. Ако в дадена квадратна област то надвишава предварително дефиниран глобален праг се счита, че тази област принадлежи към говорен фрагмент. Впоследствие чрез праговата стойност на Otsu всяко изображение 5x5 се преобразува в двоично. В него чрез експериментално дефинирани правила се елиминират звуковите артефакти, генерирани от диктора. Експерименти за ОГТ са реализирани за 15 специално подбрани думи (започващи и/или завършващи със съгласни) произнесени от 30 диктора и зашумени с адитивен шум с различно ниво. Получените резултати демонстрират предимствата на предложения подход.

1.3.2.4 Алгоритъм на Li използващ енергията на Teager

Предложен е алгоритъм за ОГТ използващ като параметър енергийния оператор на Teager (EOT) [Li, 2012]. За разлика от средно квадратичната енергия, при този вид енергия се съдържа информация не само за амплитудните, но и за честотните характеристики на сигнала. За да се определят граничните точки в контура на ЕОТ са въведени двойка прагови стойности и съответни логически правила. Тестове са реализирани с изкуствено зашумен говорен сигнал като видовете шум са от корпуса NOISEX-92 [Noisex, online]. Основния недостатък на ЕОТ при зашумен сигнал е неудовлетворителната детекция на граничните точки при наличие на беззвучни сегменти. Независимо от този факт, в сравнение със средно квадратичната енергия, получените резултати демонстрират предимствата на ЕОТ.

1.3.2.5. Алгоритъм на Roy използващ уейвлети и спектрална ентропия

В работата [Roy, 2019] се обръща особено внимание на ОГТ при наличие на не-говорни звукови артефакти, генерирани от диктора (тежко дишане и др.). Известно е, че

съществува застъпване между спектралните характеристики на артефактите и тези на редица фонеми, което води до влошаване на точността на ОГТ.

В работата е използвана Daubechies 8 уейвлет функция. Експериментите показват, че при разглежданите звукови артефакти се получават високи стойности на уейвлет коефициентите в ниските (високочестотните) скали, докато при нискочестотните скали тези коефициенти са с минимални стойности. Формирани са две множество от скали в диапазона от 300 до 3000 Hz, като броя на високочестотните скали е значително по-голям. На базата на тези множества са получени два вектора (за всеки сегмент), съответно за ниските и високите честоти и е изчислена спектралната ентропия за всеки един от тях.

Предложен е алгоритъм за ОГТ с прагови стойности изчислени отделно за всеки един вид спектрална ентропия. Експерименти са реализирани с две бази данни включващи незашумени записи на фрази от диктори в различно емоционално състояние и съдържащи значително количество от разглежданите артефакти. Постигната е точност на определяне на гранични точки: при началните 99.3% и 98.8%, а при крайните 97.5% и 99.1% - съответно за първата и втората база данни.

1.4. Заключение

Както е известно основната цел при разработване на алгоритми за детекция на говор е да се постигне робастност на тяхното решение, т.е. сегментацията на говорната последователност да не се променя независимо от промяната на качеството на сигнала и условията на средата. Освен това те трябва да имат висока точност на детекция при работа в реална среда (unconstrained environments) характеризираща се с различни по характер и често непредсказуеми шумови компоненти (unseen noise conditions).

На базата на разгледаните в обзора алгоритми може да се заключи, че за достигане на тази цел се работи в следните направления – компенсация на шума, формиране на робастни признаци, изграждане на робастни модели на класовете говор и не-говор и синтез на класификационни правила. Оказва се, че избора на признаци и класификационна схема силно зависят от особеностите на конкретната задача, което от своя страна води до ситуация, при която независимо от интензивните изследвания в областта до сега не е разработен универсален алгоритъм за детекция на говор.

Идеалният алгоритъм за детекция на говор съгласно [Li et al., 2002] трябва да отговаря на следните изисквания - надеждност, устойчивост на шум, точност, адаптивност, простота, да не изисква априорна информация за параметрите на шума, да

не изисква предварително обучени модели на класовете говор/не-говор и да работи в реално време.

При разработка на реален алгоритъм обаче се налага да се извършат редица компромиси. Ако основното изискване е постигане на висока точност при зашумени сигнали, без оглед на степента на сложност, то би трябвало да се предпочетат алгоритми, включващи методи за компенсация на шума, изчисляване на два или повече робастни признаци и използване на класификатори на базата на GMM, SVM, MLP и DNN. Тези класификатори изискват предварително обучение с учител т.е. трябва да се разполага със сегментирана говорна база данни. Ако тя не е налице се налага да се въведе или етап на предварителна груба сегментация, която да се използва като начална оценка, или да са на разположение транскрибирани данни, получени от система за автоматично разпознаване на реч. Използват се и подходи, при които се комбинират различни класификатори с цел да се преодолеят ограниченията, наложени от обучението с учител.

Ако конкретната задача изисква простота, адаптивност, липса на предварително обучение, приемлива точност, работа в реално време или с минимално закъснение, то в този случай се предпочита алгоритми с един робастен параметър, прагови стойности и логиката на краен автомат. Най-често в случая се анализира времевия контур на робастен параметър и се изчислява прагова стойност (адаптивна или фиксирана) и чрез логика на краен автомат и/или hangover схема се определя дали даден сегмент е говор, или не.

Въз основа на направения обзор може да се заключи, че комбинацията на източници доставящи различна информация е успешна стратегия при разработка на алгоритми за детекция на говор в реална среда. Тя включва комбинация на различни свойства на говорния сигнал в един признак, на различни признаци в един VAD-алгоритъм (fusion of multiple feature streams) и на различни VAD-алгоритми работещи съвместно. От своя страна съвместно работещите VAD-алгоритми могат да бъдат изградени с различни класификатори, което дава възможност за по голяма адаптивност на детектора на говор при промяна в условията на средата.

1.5. Цел на дисертацията

Целта на дисертационния труд е формиране на робастни признаци предназначени за алгоритми за детекция на говор и тяхното изследване в контекста на задачите за разпознаване на диктори при говорен сигнал записан по телефонен канал.

1.6. Задачи на дисертацията

За постигането на целта на дисертационния труд са формулирани следните основни задачи:

1. Дефиниране на робастни признаци предназначени за детекция на говор, базиращи се на свойствата на спектралната автокорелационна функция и спектъра на групово закъснение и изследване на техните характеристики.
2. Разработване на подход за определяне на граничните точки на говорно съобщение включващ алгоритъм за изчисляване на адаптивни прагови стойности и детерминиран краен автомат.
3. Разработване на алгоритми за определяне на гранични точки и експериментално изследване на тяхната ефективност с предложените в дисертацията признаци при верификация на диктори с фиксирани фрази.
4. Разработване на алгоритми за детекция на говорни сегменти и експериментално изследване на тяхната ефективност с предложените в дисертацията признаци при независима от текста идентификация на диктори.

ГЛАВА 2

Дефиниране на признаци за детекция на говор използващи свойствата на САКФ и СГЗ

2.1. Дефиниране на признаци за детекция на говор използващи спектрална автокорелационна функция

2.1.1. Увод

Едни от ефективните признаци, използвани в алгоритми за детекция на говор са тези, които се основават на хармоничната структура на говорния сигнал [Rabiner et al., 2010], [Kristjansson et al., 2005]. Известно е, че в звучните фрагменти на речта има структура, която съдържа многобройни хармоници на основния тон и която се запазва при ниско SNR. В спектралната област хармоничната структура има вид на последователност от пикове разположени на честоти кратни на основния тон.

VAD-алгоритмите използващи хармоничната структура на говора обикновено включват модул за определяне на основния тон. Приема се, че наличието на последователност от спектрални пикове с определено ниво и в определени честотни диапазони е индикатор за наличие на основен тон в даден сегмент и респективно за звучността на сегмента. При високо ниво на шума част от хармоничната структура се променя и се появяват фалшиви пикове в спектъра. За да се преодолее този факт се разработват алгоритми, при които се усилват само пиковете на хармоничните компоненти и се редуцират всички останали [Ichikawa et al., 2008], [Ishizuka et al., 2010].

В представената работа е предложено да се формират признаци за детекция на говор чрез използване на свойствата на спектралната автокорелационна функция. Основната идея е да се постигне усилване на пиковете на хармоничните компоненти в спектралната автокорелационна функция чрез използване на апроксимация на първата ѝ производна.

2.1.2. Спектрална автокорелационна функция. Свойства.

Спектралната автокорелационна функция (Spectral AutoCorrelation Function - SACF) може да бъде изчислена по различни начини съобразно целите, за които е предназначена. В работата е прието да се дефинират две версии на спектралната автокорелационна функция – съответно дефинирани чрез амплитудния спектър и чрез спектъра на

мощността. При използване на спектъра на мощността се получава автокорелационна функция със силно изразени максимуми, което е полезно в случаите, когато са налице значими шумови компоненти в сигнала. От друга страна извършените в работата изследвания показаха, че при наличие в края на думата на фрикативни или назални съгласни (low-level phonemes) по-ефективна версия на SACF (за целите в дисертацията) е тази, която се дефинира чрез амплитудния спектър.

Нека $|X(k)|$ е амплитудният спектър на говорния сигнал, получен чрез FFT за текущия сегмент. За по-голяма яснота на изложението, по-долу в текста, индекса за времеви сегмент ще бъде пропуснат. Съгласно [Klapuri, 2000] изместената оценка на автокорелационната функция $R_A(l)$ на амплитудния спектър $|X(k)|$ има вида

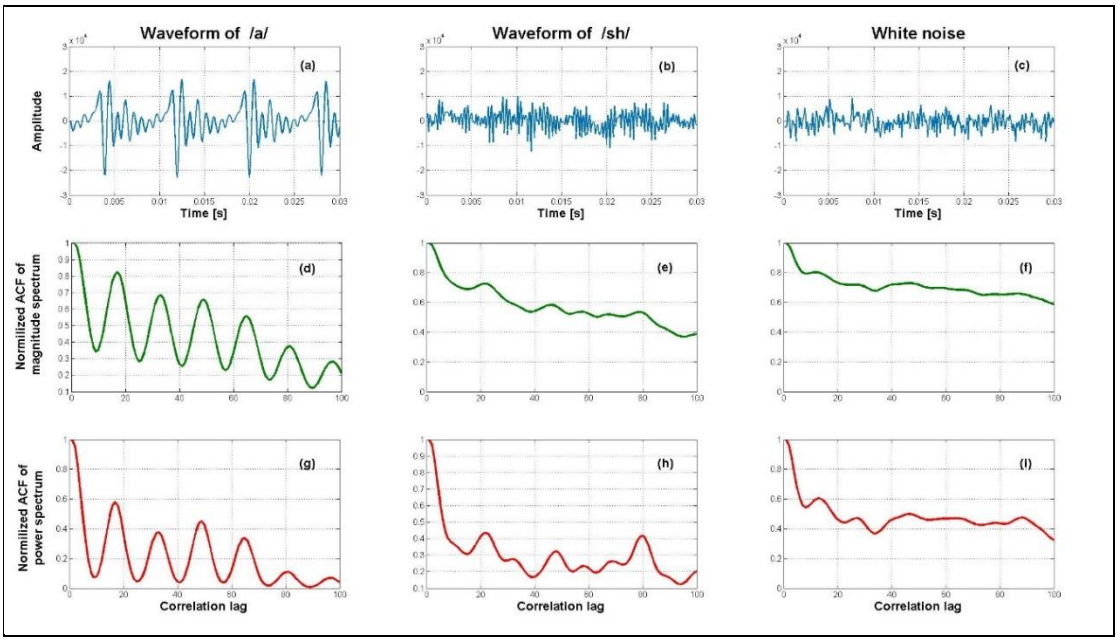
$$R_A(l) = \frac{2}{K} \sum_{k=0}^{K/2-l-1} |X(k)|.|X(k+l)|, \tag{2.1}$$

където $k = 0,...,K/2$, K е размерът на FFT, $l = 0,...,L$, L е броят на отместванията (lags).

Ако $S(k) = |X(k)|^2$ то изместената оценка на автокорелационната функция $R_P(l)$ изчислена чрез спектъра на мощността има вида

$$R_p(l) = \frac{2}{K} \sum_{k=0}^{K/2-1-l} S(k)S(k+l), \tag{2.2}$$

За по-голяма яснота по-долу в текста за SACF ще се използва означението $R(.)$ като допълнително ще бъде пояснено как точно е получена. На фиг. 2.1 са показани три сегмента с дискретизиран говорен сигнал (от корпуса BG-SRDat) и съответните нормализирани спектрални автокорелационни функции с изместване до 100 ($L = 100$). Първият сегмент съдържа част от звучната фонема /а/, вторият – част от шумовата фонема /ш/, а третия – програмно генериран бял шум. Параметрите на дискретизация и обработка са: честота на дискретизация – 8 kHz, дължина на сегмента - 240 отчета. При изчисляване на SACF дискретизирания сигнал в сегмента е умножен с тегловна функция на Хеминг. След което върху тези данни е приложено FFT с размерност 1024.



Фиг. 2.1. Нормирани спектрални автокорелационни функции съответно за (a) звучна фонема */a/*, (b) шумова фонема */ш/* и (c) бял шум.

Във всяка колона на фиг. 2.1. са показани дискретизирания сигнал и съответните нормализирани автокорелационни функции (т.е., $R_A(0)=1; R_P(0)=1$) на амплитудния и мощностния спектри. Известно е, че SACF на звучните фонеми има периодичен характер [Rabiner et al., 2010]. Освен това както се вижда на фиг. 2.1 в SACF получена чрез спектъра на мощността има много по-ясно изразени пикови стойности в сравнение с автокорелационната функция, получена чрез амплитудния спектър - виж фиг. 2.1 - (d) и (g), (e) и (h), (f) и (i)). Интересно е да се отбележи, че в SACF получена чрез спектъра на мощността се наблюдават ясно изразени пикове дори при шумовата фонема /ш/. И обратно, на фиг. 2.1 се вижда, че съществува голямо сходство между автокорелационните функции, получени чрез амплитудния спектър съответно при шумовата фонема и белия шум – виж (e) и (f).

2.1.3. Делта спектрална автокорелационна функция

В работата е предложен признак, който се основава на първата производна на спектралната автокорелационна функция. Тъй като тази производна няма аналитична форма, то тя може само да се апроксимира с крайни разлики. Обаче прилагането на крайни разлики от първи ред при реални сигнали води до увеличаване на шума, тъй като тези разлики в действителност представляват вид високочестотна филтрация. За да се избегне този проблем се предлага идея подобна на описаната в [Rabiner et al., 2010], но реализирана по различен начин. В [Rabiner et al., 2010] първата производна във времето на даден кепстрален контур е представена чрез неговата ортогонална полиномиална

апроксимация изчислена в границите на определена времева област (с продължителност от около 100-150 ms). В този случай ортогоналния полиномиален коефициент от първи ред описва наклона (т.е. първата производна във времето) на кепстралната траектория за даден времеви сегмент. За всеки кепстрален коефициент (във всеки сегмент) се извършва посочената апроксимация и се получават съответните ортогонални полиномиални коефициенти от първи ред известни под името ‘делта кепстрални коефициенти‘ или просто делта-кепстър. Делта-кепстъра за *l-я* кепстрален коефициент и *n-я* сегмент съгласно [Rabiner et al., 2010] има вида

$$\Delta d(n,l) = \frac{\sum_{k=-K}^{K} k.c(n+k,l)}{\sum_{k=-K}^{K} k^2}. \qquad (2.3)$$

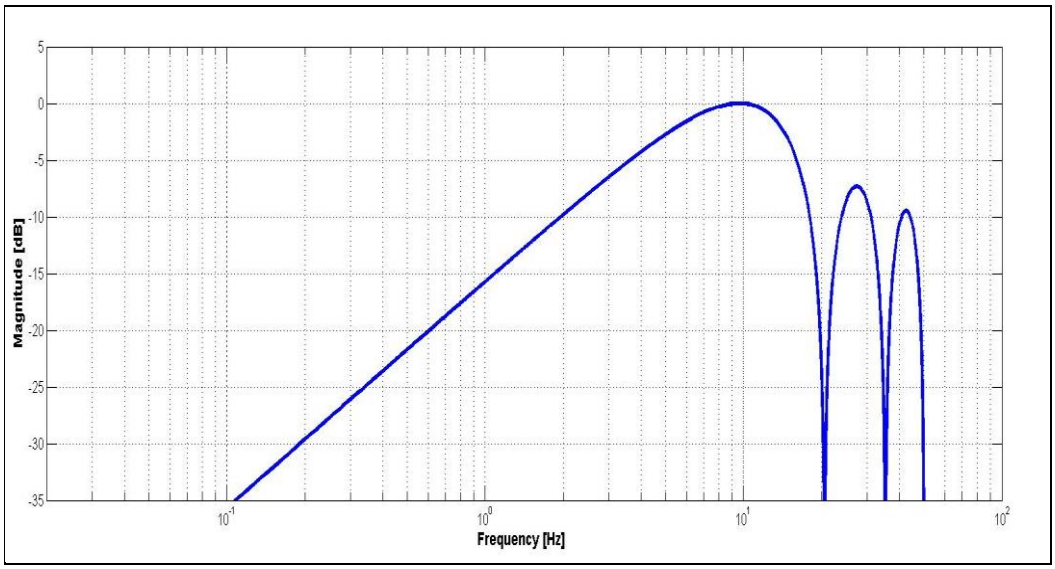
Друга интерпретация на делта-кепстъра е той да се разглежда като получен на изхода на линеен филтър с крайна импулсна характеристика [Fukuda et al., 2010]. Във времевата област делта-кепстъра $\Delta d(n,l)$ се получава от кепстралните коефициенти $c(n,l)$ и съгласно [Fukuda et al., 2010] има вида

$$\Delta d(n,l) = \frac{\sum_{k=1}^{K} k.(c(n+k,l) - c(n-k,l))}{2\sum_{k=1}^{K} k^2}, \qquad (2.4)$$

където $n = 0,...,N-1$, N е броя на сегментите; $l = 1,...,L$; L е размера на кепстралния вектор и дължината на филтъра е *2K+1*. При делта-кепстъра съгласно (2.4) се извършва филтрация на съответния времеви контур на *l-я* кепстрален коефициент. Съгласно [Fukuda et al., 2010] този филтър има предавателна характеристика $H(z)$ от вида

$$H(z) = \frac{\sum_{k=1}^{K} k.(z^k - z^{-k})}{2\sum_{k=1}^{K} k^2}. \qquad (2.5)$$

При изместване на сегментите (frame shift) 10 ms (честота на дискретизация на контура 100 Hz) и при *K=3* се усилват честотите на модулация в кепстралния контур в областта около 10 Hz [Fukuda et al., 2010]. На фиг. 2.2 е показана амплитудната честотна характеристика на филтъра (2.5) нормирана спрямо 0 dB.



Фиг. 2.2. Амплитудно честотна характеристика на КИХ филтъра от формула (2.5) (Фигурата е адаптирана версия на Фиг. 2 в [Fukuda et al., 2010]).

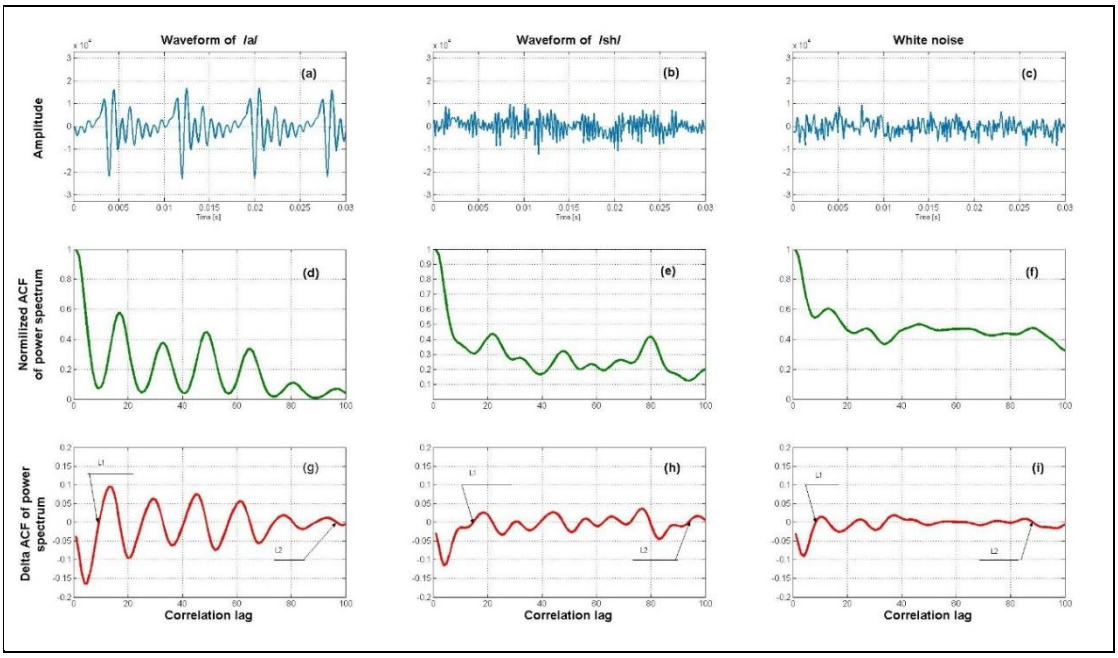
Производната на спектралната автокорелационна функция е прието да се получи като изходен сигнал на КИХ филтъра в (2.5) главно заради възможността за интерпретация на резултатите в честотната област. В работата, предложената производна е означена като Delta Spectral Autocorrelation Function (DSACF) и се изчислява, като се използват стойностите на SACF в границите само на текущия сегмент (intra-frame processing).

DSACF $\Delta R(n,l)$ за *n-я* сегмент се изчислява съгласно

$$\Delta R(n,l) = \frac{\sum_{q=1}^{Q} q.(R(n,l+q) - R(n,l-q))}{2\sum_{q=1}^{Q} q^2}, \qquad (2.6)$$

където $l = 0,...,L$ е броят измествания (lags) на SACF; Q е обикновено между 2 и 5, т.е. дължината на филтъра от 5 до 11 измествания и $n = 0,...,N-1$, N е броя на сегментите. Прието е $R(n,l) = 0$ за $l < 0$ и $l > L$, т.е. първите и последните няколко стойности на $\Delta R(n,l)$ не би трябвало да са предмет на анализ тъй като те се влияят от вида на посочените граничните условия.

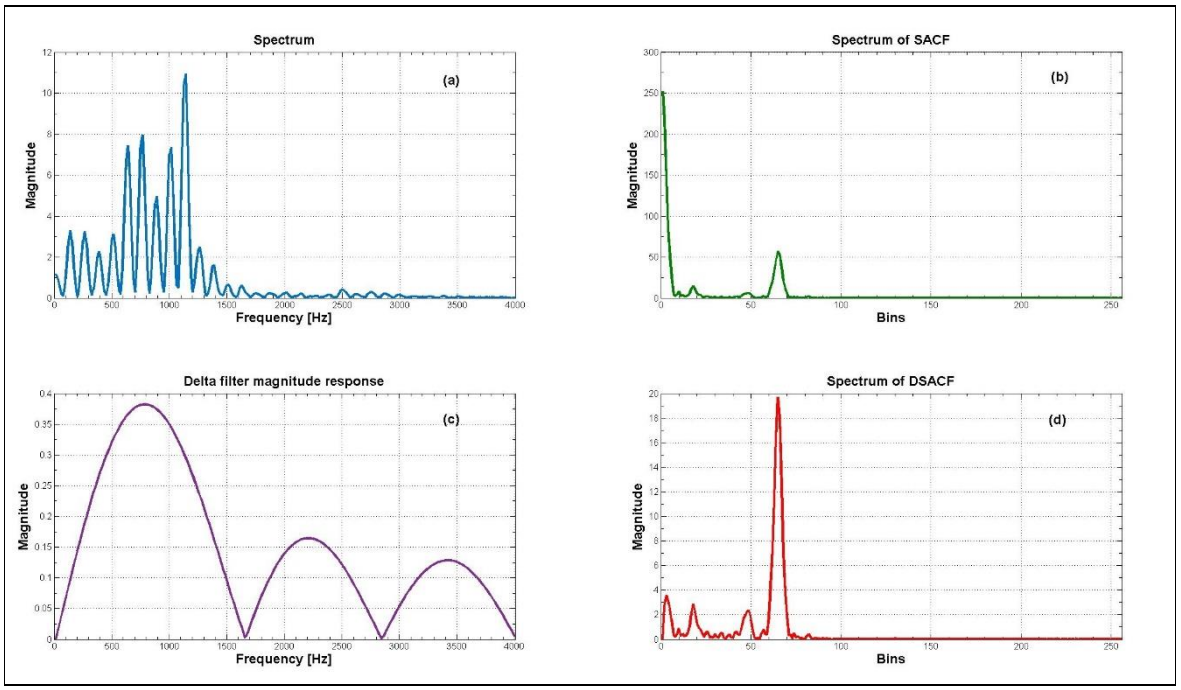
На фиг. 2.3 са показани графиките на три сигнала (същите както на фиг. 2.1), нормализираните им SACF и съответните DSACF (Q=*3*) до лаг $L = 100$.



Фиг. 2.3 Нормирани SACF и съответните DSACF за (a) звучна фонема */a/*, (b) шумова фонема */ш/* и (c) бял шум.

Спектъра на говорния сигнал за звучни фрагменти се състой от две компоненти – спектрална обвивка и хармонична структура. Обикновено разделянето им се реализира чрез кепстрален анализ [Rabiner et al., 2010]. В [Wang et al., 2001] е предложена идеята това да се осъществи чрез прилагане на второ бързо преобразуване на Фурие (2^{nd} FFT) върху вече получения амплитуден спектър. Показано е в работата, че по този начин се подчертават хармоничните пикове във втория спектър (2^{nd} FFT spectrum) съответстващи на основния тон и е значително по-лесно те да бъдат разграничени от обвивката.

При DSACF във фиг. 2.3 (g) се наблюдават силно изразени положителни и отрицателни пикови стойности, които е трудно да бъдат интерпретирани. За да се преодолее това се предлага да се използва идеята за 2^{nd} FFT, но приложено не върху амплитудния спектър, а върху спектралната и делта спектралната автокорелационни функции. По този начин е възможно да се извършва директно сравнение между двата спектъра (2^{nd} FFT spectrums) и да се установи ефекта от прилагането на филтъра (2.5) върху спектралната автокорелационна функция.



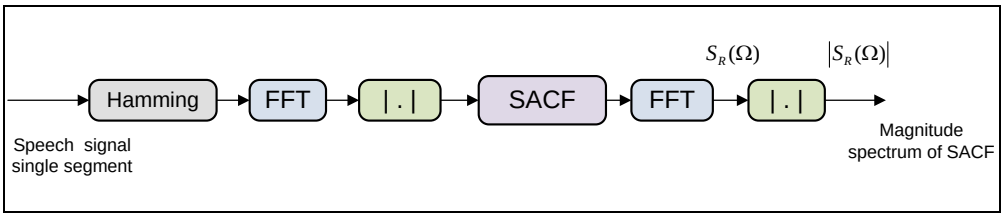
Фиг. 2.4. Амплитудни спектри на: (a) звучна фонема /a/, (b) SACF, (c) КИХ филтъра и (d) DSACF.

На фиг. 2.4 са показани съответно - амплитудния спектър на част от фонемата *'a'* (времевия ред е показан на фиг. 2.1 - честота на дискретизация 8 kHz), амплитудния спектър на SACF $|S_R(\Omega)|$, амплитудната честотна характеристика на КИХ филтъра (2.5) $|H(\Omega)|$ и амплитудния спектър на DSACF $|S_{\Delta R}(\Omega)|$. Графиките са получени при следните параметри – K=3 (формула (2.6) - дължина на филтъра 7 лага), брой точки на FFT – 512, неизместената оценка на спектралната автокорелационна функция за $n^{-я}$ сегмент по аналогия с (2.1) има вида

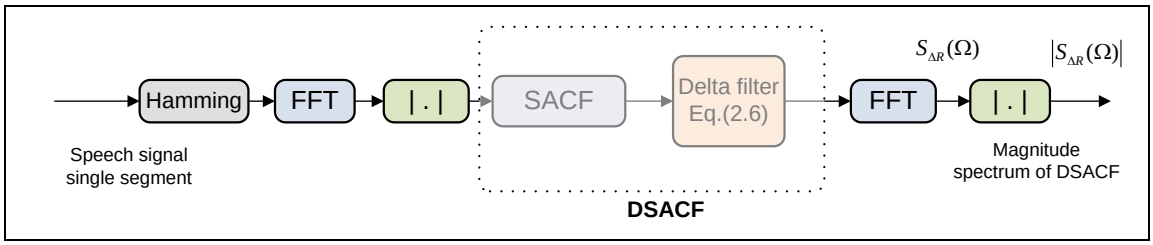
$$R(n,l)=\begin{cases}\dfrac{1}{K/2-|l|}\displaystyle\sum_{k=0}^{K/2-1-l}|X(n,k)|.|X(n,k+l)|; & l\geq 0;\ l\leq L;\\ R(n,-l) & ;l<0\end{cases}, \qquad (2.7)$$

където K е брой точки на FFT, $L=K/4$ и $|X(.)|$ е амплитудният спектър на говорния сигнал за даден сегмент, $n=0,...,N-1$, N е броя на сегментите.

Блоковите схеми на алгоритмите за изчисляване на $|S_R(\Omega)|$ и $|S_{\Delta R}(\Omega)|$ са показани на фиг. 2.5 и фиг. 2.6.



Фиг. 2.5. Блокова схема на алгоритъма за изчисляване на амплитудния спектър на SACF.



Фиг. 2.6. Блокова схема на алгоритъма за изчисляване на амплитудния спектър на DSACF.

Основния тон във фрагмента от фонемата /a/ показана на фиг. 2.4 (a) е около 125 Hz. При честота на дискретизация 8000 Hz и FFT с 512 точки разликата между пиковете на основния тон във фиг. 2.4 (a) е 8 spectrum bins (дискретни стойности на честотата в спектъра). Чрез прилагане на 2nd FFT върху SACF и DSACF и съгласно изчисленията в [Akant et al., 2010] максимума в спектъра съответстващ на основния тон е позициониран на 64 bin, както се вижда на фиг. 2.4 (b) (d).

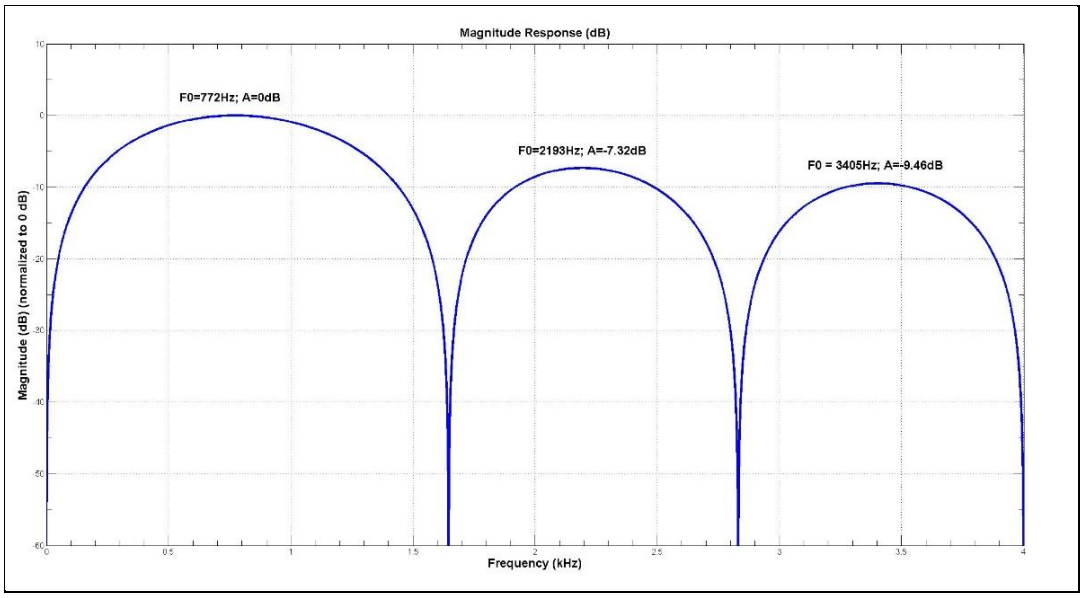
АЧХ показана на фиг. 2.4 (c) може да се разглежда условно като съвкупност от три лентови филтъра. На фигурата амплитудата е в линеен мащаб, за да е възможно сравнение с другите два амплитудни спектъра – на SACF и на DSACF. Ако амплитудата е в логаритмичен мащаб и се определят точките на ниво -3 dB спрямо максимума (0 dB) то АЧХ има вида показан на фиг.2.7 и стойностите на честотите са отразени в Таблица 2.1. С f_L и f_H са означени честотите на срязване, f_0 е централната честота и B е лентата на пропускане на съответните лентови филтри.

Таблица 2.1. Честотни параметри на делта филтъра

BPF	f_L [Hz]	f_0 [Hz]	f_H [Hz]	B[Hz]
1	383	772	1179	796
2	1904	2193	2501	597
3	3111	3405	3699	588

Първия лентов филтър има основно значение при филтриране на SACF. Амплитудно честотната му характеристика има стойност при f_0 по-голяма от тази на втория и третия филтър съответно с около 7.3 dB и 9.5 dB. Чрез този филтър се редуцират компонентите в спектъра на SACF близки до постоянната съставка (DC term) и съответстващи на енергията на обвивката на спектралната автокорелационна функция. Освен това чрез него е постигнато усилване на пика в спектъра на DSACF съответстващ на енергията на хармониците на основния тон в спектралната автокорелационна функция - както се вижда на фиг. 2.4 (d). Филтрирането на сигнала с нискочестотен филтър с гранична честота 800-1000 Hz е често използвана обработка в алгоритмите за анализ на

хармоничната структура на говорния сигнал [Rabiner et al., 2010]. Тук чрез първия лентов филтър е реализирано усилване на спектъра в диапазона 380-1100 Hz – виж таблица 2.1.



Фиг. 2.7. Логаритмична АЧХ на делта филтъра.

Спектъра $S_{\Delta R}(.)$ на DSACF е получен съгласно фиг. 2.6 чрез прилагане на FFT директно върху DSACF $\Delta R(.)$ или

$$S_{\Delta R}(.) = FFT(\Delta R(.)). \tag{2.8}$$

В резултат на въведените гранични условия при изчисляване на DSACF във формула (2.6) при получения спектър е налице леко допълнително изглаждане.

Известно е, че автокорелационната функция $r(.)$ на реален сигнал е симетрична функция [Oppenheim et al., 1999], или

$$r(-l) = r(l); \tag{2.9}$$

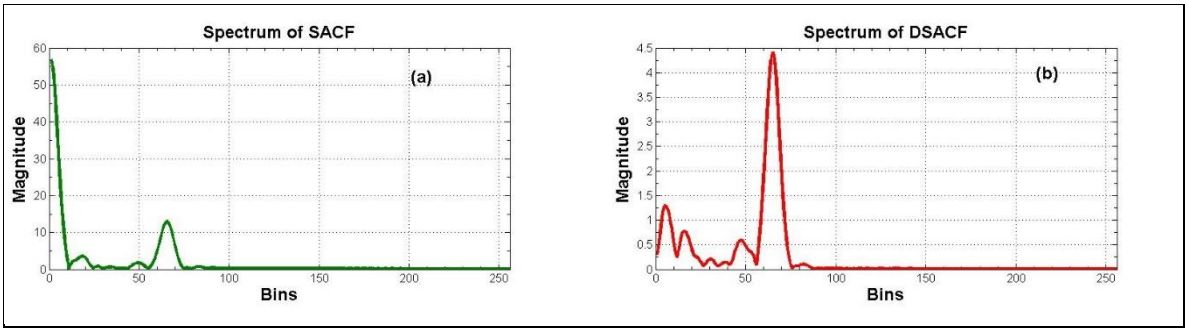
$$r(l) \le r(0); \tag{2.10}$$

Тези свойства притежават както спектралната, така и делта спектралната автокорелационни функции определени в (2.1) и (2.6). Спектрите, показани на фиг. 2.4 (b) (d) са получени след прилагане на FFT върху последователностите $(R(-L),...,R(0),...,R(L))$ и $(\Delta R(-L),...,\Delta R(0),...,\Delta R(L))$. В работата за формиране на признаци е прието да се използва само каузалната част (за положителните lags $R^p(.)$) на спектралните автокорелационни функции, а именно

$$R^p(l)=\left\{\begin{array}{ll} R(l); & l>0 \\ R(0)/2; & l=0; \\ 0; & l<0; \end{array}\right\};\ R(l)=R^p(l)+R^n(l);\quad -L\le l\le L; \tag{2.11}$$

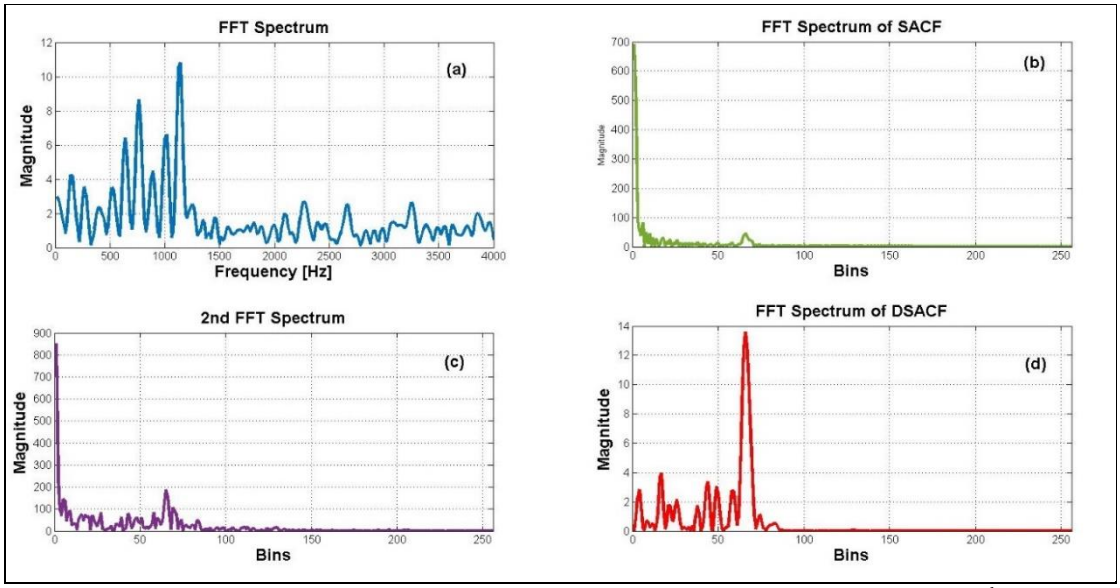
$$\Delta R^p(l)=\left\{\begin{array}{ll} \Delta R(l); & l>0 \\ \Delta R(0)/2; & l=0; \\ 0; & l<0; \end{array}\right\};\ \Delta R(l)=\Delta R^p(l)+\Delta R^n(l);\quad -L\le l\le L;, \tag{2.12}$$

където $R^n(.)$ и $\Delta R^n(.)$ са автокорелационните функции за отрицателни lags. На фиг. 2.8 са показани амплитудните спектри на $R^p(.)$ и $\Delta R^p(.)$ за сегмента на фонемата /a/ от фиг. 2.3(a). С цел да се редуцират стойностите на автокорелационната функция с lags близък до 0 тя се умножава с тегловна функция на Hamming.



Фиг. 2.8. Амплитудни спектри на: (a) SACF и (b) DSACF

На фиг. 2.9 са показани описаните по-горе FFT амплитудни спектри, но изчислени за говорен сигнал с добавен бял шум с Гаусово разпределение и SNR=5 dB.



Фиг. 2.9. SNR=5 dB - Амплитудни спектри на: (a) звучна фонема /a/, (b) SACF, (c) 2nd FFT и (d) DSACF.

Ако се сравнят спектрите при чист и зашумен сигнал показани съответно на фиг. 2.4 и фиг. 2.9 ще се установи следното. Първо, потвърждава се идеята на [Wang et al., 2001] за подчертаване на пика на основния тон при зашумени сигнали. На фиг. 2.9 (c) пикът в 2nd FFT спектъра, разположен на 64 bin съответства на основния тон от 125 Hz. Второ, при сравнение на спектрите съответно на SACF – фиг. 2.4 (b) и 2.9 (b) и на DSACF – фиг. 2.4 (d) и 2.9 (d) се установява, че пика в спектъра на SACF е редуциран в значително по-голяма степен, отколкото съответния пик в спектъра на делта спектралната автокорелационна функция. Освен това пика в DSACF за зашумен сигнал е по-силно изразен дори от този в 2nd FFT спектъра. Тези факти са аргументи за използване на DSACF като основа за формиране на робастни признаци за детекция на говор.

2.1.4. Среден-делта признак - Mean-Delta (MD) feature

2.1.4-A. Мотивация

Описаните в т. 2.1.3 характеристики на DSACF са основа на предложените в дисертацията признаци. Както се вижда на фиг. 2.3 (g) (h) (e) DSACF притежава значими положителни и отрицателни пикове дори за фрикативната съгласна ‚*ш*'. Това свойство от една страна, а от друга, формата на спектъра на DSACF за зашумени сигнали, показана на фиг. 2.9 (d), са отправни точки при формиране на предложените в дисертацията признаци. Авторът предполага, че ако се формира параметър, който за текущия сегмент да представлява сумарна оценка на броя и големината на пиковете в DSACF то този параметър може да се използва успешно като признак за детекция на говор особено при зашумени сигнали. В дисертацията това предположение е потвърдена експериментално за две версии на DASCF изчислени съответно чрез амплитудния спектър на Фурие и чрез модифицирания спектър на групово закъснение.

Предложените в Глава 2 признаци за детекция на говор са формирани чрез DSACF, а не чрез нейния спектър. Директното използване на спектъра на DSACF (т.е. прилагане на второ FFT) за формиране на признаци предназначени за детекция на говор и проверката на тяхната ефективност в системи за разпознаване на диктори е предмет на бъдещи изследвания.

2.1.4.1. Среден-Делта признак

Първия предложен признак е наречен среден-делта признак (Mean-Delta - MD feature) и е предназначен за използване при анализ на времеви контури. За $n^{-я}$ сегмент MD признака $m_d(n)$ е дефиниран съгласно

$$m_d(n) = F\left(\sum_{l=0}^{L}\left|\Delta R(n,l)\right|\right), \tag{2.13}$$

където $\Delta R(n,l)$ е каузалната част на DSACF изчислена съгласно (2.12). Във формула (2.13) с *F(.)* е означено допълнително преобразуване което е дефинирано съобразно особеностите на алгоритъма за детекция на говор. Тук ще бъде представен т.н. наречения *основен* алгоритъм за определяне на MD признака. За $n^{-я}$ сегмент той има вида:

- за претегления с тегловна функция на Хеминг говорен сигнал се изчислява амплитудния спектър $\left|X(k)\right|$ чрез FFT с размер *K*;
- извършва се нормализация спрямо средния вектор на амплитудния спектър (изчислен по всички сегменти в текущото произнасяне)

$$\left|\hat{X}(n,k)\right| = \frac{\left|X(n,k)\right|}{\frac{1}{N}\sum_{n=1}^{N}\left|X(n,k)\right|}, \tag{2.14}$$

  където *N* е броят на сегментите в произнасянето;
- определя се неизместената оценка на спектралната автокорелационна функция с lags *L=K/4*, като се използва нормализирания амплитуден спектър

$$R(n,l) = \frac{1}{K/2-|l|}\sum_{k=0}^{K/2-1-l}\left|\hat{X}(n,k)\right|.\left|\hat{X}(n,k+l)\right|; \quad l \geq 0;\ l \leq L; \tag{2.15}$$

- определя се делта спектралната автокорелационна функция $\Delta R(n,l)$ съгласно (2.6) с *Q=3*;
- изглажда се контура във времето на делта спектралната автокорелационна функция (за всеки лаг) чрез Long-Term Spectral Envelope (LTSE) algorithm с параметър *J=3* [Ramirez et al., 2004]. Така получената изгладена версия на $\Delta R(n,l)$ е означена като $\Delta R^s(n,l)$

$$\Delta R^s(n,l) = \max\left\{\Delta R(n+j,l)\right\}_{j=-J}^{j=+J}. \tag{2.16}$$

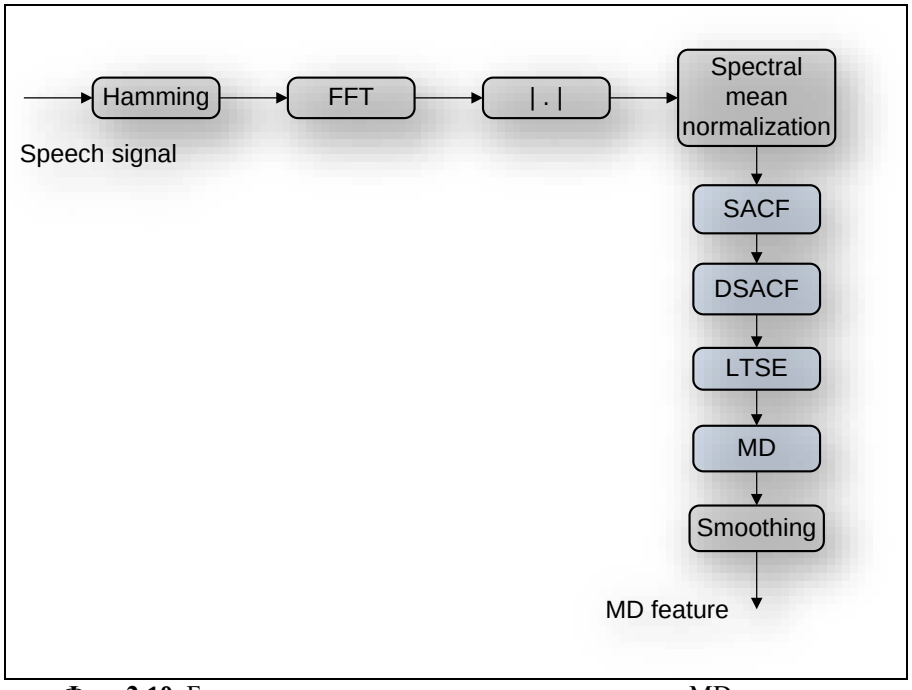
- изчислява се MD признака $m_d(n)$

$$m_d(n) = \left[\sum_{l=0}^{L}\left|\Delta R^s(n,l)\right|\right]^{0.5} \tag{2.17}$$

- изглаждане на $m_d(n)$ контура чрез филтър с изместваща се средна стойност;

На фиг. 2.10 е показана блок схемата на описания по-горе алгоритъм за определяне на MD признака.

Описания алгоритъм за определяне на MD-признака условно е назован по-горе в текста *основен*, защото в някой от публикациите на автора са налице изследвания, при които са въведени несъществени модификации в алгоритъма, например – вместо изглаждане с LTSE алгоритъм се използва филтър с изместваща се средна стойност и др. Ако е необходимо при описание на проведените експерименти в Глава 3 и Глава 4 наличието на тези модификации ще бъдат указано.



Фиг. 2.10. Блок схема на алгоритъма за определяне на MD признака.

2.1.4.2. Базов среден-делта признак

Втория признак носи името базов среден-делта признак (Basic Mean-Delta (BMD) feature) и е предназначен за детекция на говор в алгоритми за разпознаване, т.е. параметъра е дефиниран във векторна форма. За $n^{-я}$ сегмент BMD признака $m_{BMD}(n)$ се определя по следния начин:

- за претеглените с тегловна функция на Хеминг говорен сигнал се изчислява амплитудния спектър $|X(k)|$ чрез FFT с размер K;
- определя се амплитудния спектър $|\hat{X}(n,k)|$ съгласно (2.14) нормиран спрямо средната си стойност (изчислена по всички сегменти в текущото произнасяне);
- определя се неизместената оценка съгласно (2.15) на спектралната автокорелационна функция с lags $L=K/4$, като се използва нормализирания амплитуден спектър;
- определя се делта спектралната автокорелационна функция съгласно (2.6) с $Q=3$;

- изглажда се контура във времето на делта спектралната автокорелационна функция (за всеки лаг) чрез Long-Term Spectral Envelope (LTSE) algorithm с параметър *J=3* [Ramirez et al., 2004]. Така получената изгладена версия на $\Delta R(n,l)$ е означена като $\Delta R^s(n,l)$

$$\Delta R^s(n,l) = \max\left\{\Delta R(n+j,l)\right\}_{j=-J}^{j=+J}. \tag{2.18}$$

- общия брой изместванията (lags) *L* при DSACF се разделя на *V* равни по дължина и незастъпващи се диапазони във вида

$$\{L_1, L_2\}...\{L_v, L_{v+1}\}...\{L_{2V-1}, L_{2V}\} \tag{2.19}$$

- броят диапазони *V* определя размера на BMD вектора $m_{BMD}(n)$ или той има вида

$$m_{BMD}(n) = \{m_{BMD}(n,1),...,m_{BMD}(n,v),...,m_{BMD}(n,V)\} \tag{2.20}$$

- $v^{-я}$ компонент на $m_{BMD}(n,v)$ се определя съгласно

$$m_{BMD}(n,v) = \log\left[\max\left\{\left|\Delta R^s(n,m)\right|\right\}_{m=L_v}^{m=L_{v+1}}\right] \tag{2.21}$$

2.1.4.3. Модифициран среден-делта признак

Третия признак е наречен модифициран среден-делта признак (Modified Mean-Delta (MMD) feature) и е предназначен за детекция на говор чрез алгоритми за разпознаване. Той се дефинира по начин подобен на базовия MD признак от подточка 2.1.4.2. Разликата е, че в последователността от lags при DSACF е дефинирана тегловна функция с правоъгълна форма с дължина *Y*, която се измества със стъпка *U* lags и броят стъпки *V* определя размера на MMD вектора или $m_{MMD}(n)$ за $n^{-я}$ сегмент е

$$m_{MMD}(n) = \{m_{MMD}(n,1),...,m_{MMD}(n,v),...,m_{MMD}(n,V)\} \tag{2.22}$$

където $m_{MMD}(n,v)$ има вида

$$m_{MMD}(n,v) = \log\left[\max\left\{\left|\Delta R^s(n,m)\right|\right\}_{m=(v-1)*U}^{m=(v-1)*U+Y}\right] \tag{2.23}$$

2.2. Дефиниране на признаци за детекция на говор използващи спектър на групово закъснение

2.2.1. Увод

При детекция на говор най-често се използват параметрични представяния основаващи се на свойствата на кратковременния спектър. Това е така, защото фазовия спектър притежава шумоподобен характер. Фазовият спектър обаче притежава и свойства, които могат да бъдат полезни на етапа на предварителната обработка на говорни сигнали. Те

се проявяват в спектъра получен като отрицателна първа производна спрямо честотата на фазовия спектър и наречен спектър на групово време на закъснение или Спектър на Групово Закъснение (СГЗ) - (Group Delay Spectrum - GDS) [Murthy et al., 2011].

В тази подточка са разгледани основните методи за определяне на GDS и е извършен качествен анализ на изменението на GDS при зашумени с адитивен шум говорни сигнали. Този анализ е реализиран косвено, чрез изследване на изменението на аргументите на Проекционните Функции на Сходство (ПФС) на основата на адитивния спектрален модел [Mansour et al., 1989]. Предложено е параметрично представяне наречено Group Delay Mean Delta (GDMD) признак съчетаващо модифицирания СГЗ (Modified GDS - MGDS) [Hegde et al., 2007] и MD признака, разгледан в т.2.1.4.

2.2.2. Спектър на групово закъснение

Нека $X(\Omega) = |X(\Omega)|\exp(j\Theta(\Omega))$ е преобразуване на Фурие на дискретния сигнал $\{x(l)\}$. Спектърът на групово закъснение $\tau(\Omega)$ се определя като отрицателна първа производна спрямо честотата Ω на фазовия спектър [Murthy et al., 2011]

$$\tau(\Omega) = -\frac{d\Theta(\Omega)}{d\Omega}, \tag{2.24}$$

където $\Theta(\Omega)$ е фазовият спектър в непрекъсната форма

$$\Theta(\Omega) = \Theta_v(\Omega) + \pi\sigma(\Omega), \tag{2.25}$$

$\Theta_v(\Omega)$ е главната стойност на фазата получена чрез функцията *arctan*, а $\sigma(\Omega)$ е целочислена функция, която указва колко π трябва да се добави към $\Theta_v(\Omega)$, за да се получи непрекъсната функция на фазата. Определянето на фазовата функция в непрекъснат вид при реални сигнали е съпътствано със значителни затруднения, особено при непосредствена близост на нулите на z-преобразуването на сигнала до единичната окръжност [Murthy et al., 2011].

2.2.2.1. Определяне на GDS чрез производна на логаритмична функция.

Ако се предположи [Murthy et al., 2011], че $Y(\Omega) = \ln(X(\Omega))$ е диференцируема комплексна функция и се разгледа първата й производна $Y^{\wedge}(\Omega)$ спрямо Ω върху единичната окръжност, то

$$Y^{\wedge}(\Omega) = \frac{1}{X(\Omega)} dX(\Omega)/d\Omega = \mathrm{Re}\left[Y^{\wedge}(\Omega)\right] + j\,\mathrm{Im}\left[Y^{\wedge}(\Omega)\right], \tag{2.26}$$

където $j = \sqrt{-1}$ и Re[] е реална част, а Im[] - имагинерна част на израза в скобите.

Като се има предвид, че $\ln(X(\Omega)) = \ln\left(X(\Omega)\right(+ j\Theta(\Omega)$, то от (2.26) следва, че първата производна на фазата $\Theta(\Omega)$ спрямо Ω е

$$\Theta^{\wedge}(\Omega) = \mathrm{Im}\left[Y^{\wedge}(\Omega)\right]. \tag{2.27}$$

От (2.26) и (2.27) следва, че спектърът на групово закъснение е

$$\tau(\Omega) = -\Theta^{\wedge}(\Omega) = -\frac{\mathrm{Re}\left[X(\Omega)\right].\mathrm{Im}\left[X^{\wedge}(\Omega)\right] - \mathrm{Re}\left[X^{\wedge}(\Omega)\right].\mathrm{Im}\left[X(\Omega)\right]}{\mathrm{Re}\left[X(\Omega)\right]^2 + \mathrm{Im}\left[X(\Omega)\right]^2}, \tag{2.28}$$

където $X^{\wedge}(\Omega)$ е първата производна спрямо Ω на трансформацията на Фурие на времевата последователност $\{x(l)\}$ и се определя като

$$X(\Omega) = \sum_{l=0}^{\infty} x(l)\exp(-j\Omega l); \tag{2.29}$$

$$X^{\wedge}(\Omega) = \frac{dX(\Omega)}{d\Omega} = -j\sum_{l=0}^{\infty} lx(l)\exp(-j\Omega l); \tag{2.30}$$

т.е $\mathrm{Re}\left[X^{\wedge}(\Omega)\right] + j\,\mathrm{Im}\left[X^{\wedge}(\Omega)\right] = -j.FT\left\{lx(l)\right\}$, където *FT* е преобразуване на Фурие.

При този метод се получават точните стойности на спектъра на групово закъснение $\tau(\Omega)$, без да се налага непосредственото диференциране на фазовия спектър, но е необходимо двукратно прилагане на преобразуването на Фурие.

2.2.2.2. Определяне на GDS чрез кепстрални коефициенти

При минимално-фазов сигнал $\ln\left|X(\Omega)\right|$ и непрекъснатата фаза $\Theta(\Omega)$ са свързани чрез кепстралните коефициенти [Murthy et al., 2011]. Ако $\{x(l)\}$ е минимално-фазова последователност, $X(\Omega) = \left|X(\Omega)\right|e^{j\Theta(\Omega)}$ и $\{c(n)\}$ е кепстъра на $\{x(l)\}$, то $\ln\left(X(\Omega)\right)$ може да се представи във вида

$$\ln(X(\Omega)) = \frac{c(0)}{2} + \sum_{n=1}^{\infty} c(n)\exp(-j\Omega n). \tag{2.31}$$

Тогава

$$\ln\left|X(\Omega)\right| = \frac{c(0)}{2} + \sum_{n=1}^{\infty} c(n)\cos(\Omega n) \tag{2.32}$$

и $\Theta(\Omega)$ се представя като

$$\Theta(\Omega) = -\sum_{n=1}^{\infty} c(n)\sin(n\Omega). \tag{2.33}$$

В общия случай се дефинират два спектъра на групово закъснение - от фазата $\tau_p(\Omega)$ и от модула $\tau_m(\Omega)$. При минимално-фазов сигнал те са равни и $\tau_p(\Omega) = \tau_m(\Omega) = \tau(\Omega)$

$$\tau(\Omega) = -\frac{d\Theta(\Omega)}{d\Omega} = \sum_{n=1}^{\infty} nc(n)\cos(\Omega n). \quad (2.34)$$

При смесено-фазов сигнал се въвеждат два вида кепстрални коефициенти $\{c_m(n)\}$ и $\{c_p(n)\}$, получени от минимално-фазовите еквиваленти на сигнала $\{x(l)\}$ съответно от модула и от фазата на преобразуването на Фурие.

$\tau_p(\Omega)$ и $\tau_m(\Omega)$ се определят като

$$\tau_p(\Omega) = \sum_{n=1}^{\infty} nc_p(n)\cos(\Omega n); \quad (2.35)$$

$$\tau_m(\Omega) = \sum_{n=1}^{\infty} nc_m(n)\cos(\Omega n). \quad (2.36)$$

GDS притежава две основни свойства [Murthy et al., 2011]: адитивност - следва от адитивността на фазата и висока разрешаваща способност - значимата информация за даден резонатор е концентрирана около резонансната му честота в спектъра.

2.2.3. Изследване на GDS при зашумени с адитивен шум говорни сигнали

При детекция на зашумени [Krishnan et al., 2006], [Padmanabhan et al., 2008] и синтетични [Sahidullah et al., 2015], [Tian et al., 2015] говорни сигнали са използвани признаци основаващи се на свойствата на GDS и резултатите са по-добри от тези получени чрез различни кепстрални представяния. Една от възможните причини е, че в сравнение с кепстъра GDS се влияе по-слабо от адитивен шум. За да се анализира това предположение в тази подточка е извършен качествен анализ на поведението на GDS при зашумени с адитивен шум говорни сигнали.

2.2.3.1. Общи положения

Характерът на изменение на кепстъра получен чрез метода на линейно предсказване (ЛП-кепстър) при зашумени с адитивен шум говорни сигнали е изследван теоретично и експериментално в [Mansour et al., 1989]. Изследванията са реализирани косвено, чрез анализ на изменението на аргументите на ПФС. Тези аргументи са коефициентът на редукция на нормата на вектора на зашумения сигнал и функцията на ъгъла между векторите на чистия и зашумения сигнал. Установено е, че при увеличаване нивото на адитивния шум, нормата на вектора съставен от кепстралните коефициенти на

зашумения сигнал се редуцира, а ъгълът между кепстралните вектори на чистия и зашумения сигнал се увеличава.

В контекста на изложеното в [Mansour et al., 1989] е и предметът на настоящата работа. В нея е изследвано (по подобен начин) изменението на GDS при зашумени с адитивен шум говорни сигнали и резултатите са сравнени с тези за ЛП-кепстъра.

2.2.3.2. Спектър на групово закъснение

Ако *c(n)* е ЛП-кепстъра на анализирания говорен сигнал, то спектърът на групово закъснение $\tau(\Omega)$ съгласно (2.34) е

$$\tau(\Omega) = \sum_{n=1}^{\infty} w(n)c(n)\cos(n\Omega) = \sum_{n=1}^{\infty} nc(n)\cos(n\Omega), \qquad (2.37)$$

където *w(n)* е тегловна функция и *w(n) = n*.

СГЗ определен съгласно (2.37) се отличава със силно изразени пикове. Това свойство налага въвеждане на допълнително изглаждане, което се реализира чрез подходящ избор на тегловна функция *w(n)* и/или чрез ограничаване на броят на членовете в сумата, само до първите няколко кепстрални коефициента. В [Singer et al., 1990] за изглаждане на GDS е използвана синусоидална тегловна функция (СТФ) $w_s(n)$, която има вида

$$w_s(n) = \sin(nB)/B; \qquad (2.38)$$

Експерименталното изследване в работата е при зашумени с адитивен шум говорни сигнали и има две цели: първо, да провери верността на теоретичните изводи и второ, да сравни (чрез хистограми) изменението на аргументите на ПФС, изчислени съответно за изгладения със СТФ GDS $\tau_g(\Omega)$ (определен в (2.37) при $w(n) = w_s(n)$) и за ЛП-кепстъра.

2.2.3.3. Изследване на GDS при зашумени с адитивен шум говорни сигнали

Формално задачата за изследване на изменението на GDS, при зашумени с адитивен шум говорни сигнали се решава, като се определят статистическите свойства на зашумения спектър, при положение че са известни статистическите свойства на говорния сигнал и на адитивния шум. Определянето на GDS се извършва чрез кепстрални коефициенти, което налага въвеждане на нелинейни операции. Подобни операции водят до значителни затруднения при определяне на статистическите свойства на зашумения спектър. Поради тази причина, изследванията в работата са реализирани косвено [Mansour et al., 1989].

2.2.3.4. Анализ чрез адитивен спектрален модел

Аргументите на ПФС са: коефициент на редукция на нормата на зашумения вектор - $\alpha_c = \|c_t\|/\|c_r\|$ и функция на ъгъла δ_c между два вектора - $\beta_c = 1-\cos(\delta_c) = 1-c_t.c_r/(\|c_t\|.\|c_r\|)$. c_r и c_t са вектори от коефициентите на ЛП-кепстъра $\{c_r(n)\}$ и $\{c_t(n)\}$, съответно c_r за чистия и c_t за зашумения сигнал, $c_t.c_r$ е скаларното им произведение, а $\|.\|$ - норма на вектор. За да се сравнят резултатите, получени в [Mansour et al., 1989] с тези за $\tau(\Omega)$, в настоящата работа са разгледани същите аргументи, но изчислени за $\tau(\Omega)$. Те имат вида: $\alpha_\tau = \|\tau_t\|/\|\tau_r\|$ и $\beta_\tau = 1-\cos(\delta_\tau) = 1-(\tau_t.\tau_r)/(\|\tau_r\|.\|\tau_t\|)$, където τ_r и τ_t са вектори от коефициенти на $\tau(\Omega)$ съответно за чистия и за зашумения сигнал. Изменението на посочените аргументи при GDS е изследвано косвено (чрез анализ на претеглията с индекса ЛП-кепстър *{nc(n)}*) с помощта на адитивен спектрален модел. Подобно косвено изследване е допустимо, тъй като, от една страна, нормата и ъгъла при GDS са непосредствено свързани с $\{nc(n)\}$ чрез съотношението на Парсевал [Oppenheim et al., 1999], а от друга, тук интерес представлява само характера на изменение на тези аргументи.

Съгласно адитивния спектрален модел, мощностният спектър $X(\Omega)$ на зашумения говорен сигнал е $X(\Omega) = 1/|A(\Omega)|^2 + \Phi$, където Φ - мощностен спектър на шума (мощността е константа и чрез нея се определя SNR) и $1/|A(\Omega)|^2$ - мощностен спектър на чистия говорен сигнал, определен чрез метода на ЛП [Mansour et al., 1989].

Въведено е означението $X^{\wedge}(\Omega) = dX(\Omega)/d\Omega$ (при $d\Phi/d\Omega = 0$) и съгласно [Oppenheim et al., 1999] $\{nc_t(n)\}$ е

$$nc_t(n) = -\frac{1}{j2\pi}\int_{-\pi}^{\pi}\left(X^{\wedge}(\Omega)/X(\Omega)\right)\exp(j\Omega n)d\Omega, \qquad (2.39)$$

където $n > 0$, $j = \sqrt{-1}$.

Енергията $E(\Phi)$ на $\{nc_t(n)\}$ и спектъра $jX^{\wedge}(\Omega)/X(\Omega)$ са свързани чрез съотношението на Парсевал във вида

$$E(\Phi) = 2\sum_{n=1}^{\infty}[nc_t(n)]^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{jX^{\wedge}(\Omega)}{X(\Omega)}.\frac{-jX^{\wedge}(\Omega)}{X(\Omega)}d\Omega. \qquad (2.40)$$

След заместване на $X(\Omega)$ и $X^{\wedge}(\Omega)$ в (2.40) и като се има предвид, че $dX^{\wedge}(\Omega)/d\Phi = 0$, то за $dE(\Phi)/d\Phi$ се получава

$$\frac{dE(\Phi)}{d\Phi} = \frac{1}{\pi}\int_{-\pi}^{\pi} \frac{-4\left(d\left|A(\Omega)\right|/d\Omega\right)^2}{\left(1+\Phi\left|A(\Omega)\right|^2\right)^3} d\Omega \qquad (2.41)$$

От (2.41) следва, че $dE(\Phi)/d\Phi \le 0$. Максималната стойност на $E(\Phi)$ е при $\Phi = 0$ и $E(\Phi)$ намалява при нарастване на Ф, т.е. нормата на $\{nc_t(n)\}$ намалява при увеличаване нивото на адитивния шум и $\alpha_\tau \le 1$.

Съгласно (2.39) и съотношението на Парсевал и като се има предвид, че за чист сигнал $\Phi = 0$, скаларното произведение на $\{nc_t(n)\}$ и $\{nc_r(n)\}$ е

$$2\sum_{n=1}^{\infty} nc_t(n)nc_r(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{4\left(d\left|\mathrm{A}(\Omega)\right|/d\Omega\right)^2}{\left|A(\Omega)\right|^2\left(1+\Phi\left|A(\Omega)\right|^2\right)} d\Omega \qquad (2.42)$$

От (2.42) следва, че скаларното произведение е винаги неотрицателно ($\beta_\tau \le 1$) и β_τ расте при увеличаване на Ф.

Направеният анализ показва, че при повишаване на нивото на адитивния шум (т.е. Ф расте) се наблюдава: първо, редуциране на нормата при $\{nc_t(n)\}$ и увеличаване на ъгъла между $\{nc_t(n)\}$ и $\{nc_r(n)\}$ и второ, диапазоните на изменение на α_τ и β_τ при *{nc(n)}* са аналогични на тези при *{c(n)}*. Съгласно (2.37) тези изводи за в сила и за GDS $\tau(\Omega)$.

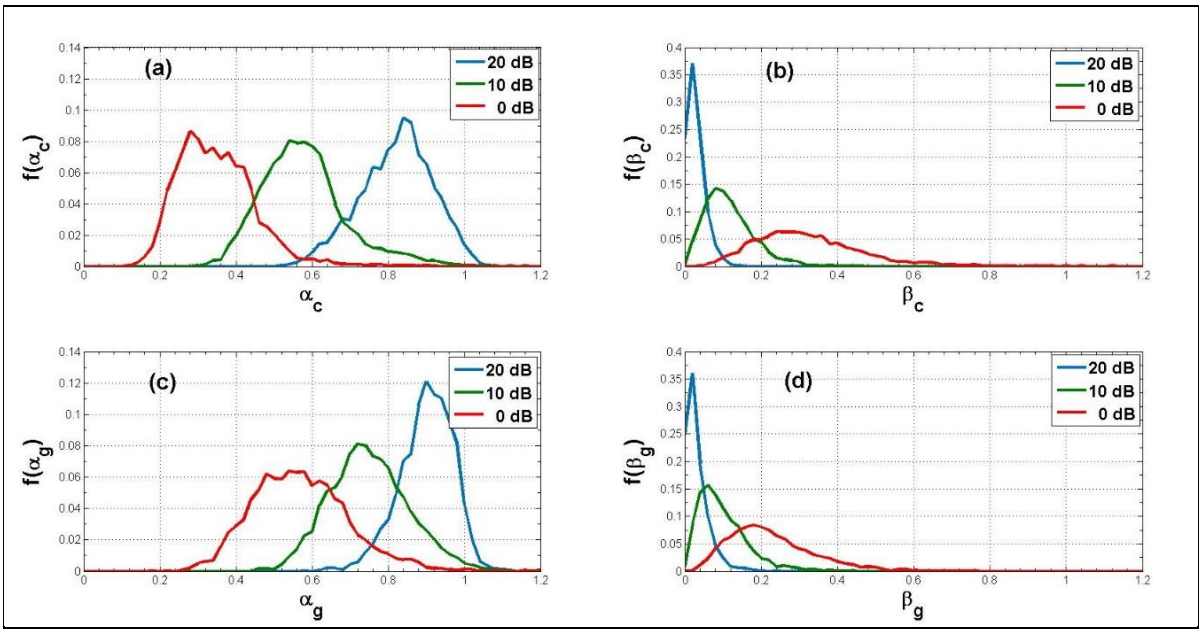
Необходимо е, да се има предвид, че направените изводи се отнасят до вектора *{nc(n)}* съдържащ всички елементи, в случая *n=1,2,...,∞*. При положение че се използват само част от елементите на вектора, т.е. *n=1,2,...,M*, при *M* << ∞ (както е в действителност при извършените експерименти) е възможно да се получат стойности на отношението α надвишаващи незначително 1.0 или скаларното произведение в (2.42) да приеме отрицателни стойности.

2.2.3.5. Анализ чрез хистограми

Използваните говорни данни са запис чрез микрофон на прочетен текст от диктор мъж. Дължината на записа е около 41 сек. (4198 сегмента). Честотата на дискретизация е 8 kHz, дължината на сегмента е 30 ms, а изместването 10 ms. Използван е автокорелационен метод за ЛП при ред на модела - 12 и брой кепстрални коефициенти - 12. Параметъра на $w_s(n)$ е $B = \pi/32$ [Singer et al., 1990]. Зашуменият говорен сигнал е получен чрез добавяне към чистия сигнал на шум с нормално разпределение и с нулева средна стойност. Изследванията са реализирани при SNR 20, 10 и 0 dB. На Фиг. 2.11 (а)

(b) са представени хистограми на α_c и β_c за ЛП-кепстъра, а на Фиг. 2.11 (c)(d) - хистограми на α_g и β_g за GDS, изгладен чрез СТФ и с допълнително редуцирани отрицателни стойности в спектъра [Zhu et al., 2004].

Въведени са коефициенти на редукция γ и η, съответно $\gamma = \bar{\alpha}[AdB]/\bar{\alpha}[20dB]$ и $\eta = \bar{\beta}[AdB]/\bar{\beta}[20dB]$, където A=20, 10 и 0 dB. В Таблица 2.2 са показани средните стойности на α_c, α_g и β_c, β_g , както и стойностите на $\gamma_c, \gamma_g, \eta_c$ и η_g при SNR 20,10 и 0 dB.



Фиг. 2.11. Хистограми при различни SNR (20,10 и 0 dB) на параметрите: (a) ЛП-кепстър α_c ; (b) ЛП-кепстър β_c; (c) GDS α_g и (d) GDS β_g .

Таблица 2.2. Средни стойности

Параметри	**20 dB**	**10 dB**	**0 dB**
$\bar{\alpha}_c$	0.8209	0.5780	0.3487
$\bar{\alpha}_g$	0.9006	0.7496	0.5694
$\bar{\beta}_c$	0.0284	0.1154	0.3211
$\bar{\beta}_g$	0.0311	0.0986	0.2249
γ_c	1	0.7040	0.4248
γ_g	1	0.8323	0.6322
η_c	1	4.0541	11.2833
η_g	1	3.1616	7.2119

2.2.3.6. Дискусия и изводи

Експерименталните резултати, представени на фиг. 2.11 и таблица 2.2 потвърждават изводите в т. 2.2.3.4, а именно, че адитивния шум предизвиква редукция на нормата и увеличаване на ъгъла при GDS като основния диапазон на изменение на α_g и β_g съвпада с този при α_c и β_c.

Сравнението между хистограмите на фиг. 2.11 (a) и (c), както и на стойностите на коефициентите на редукция γ_c и γ_g в таблица 2.2 показва, че при едно и също ниво на адитивен шум, средната стойност на α_g е по-голяма от α_c, т.е. редукцията на нормата при GDS $\tau_g(\Omega)$, е по-малка от тази при ЛП-кепстъра (това е по-силно изразено при ниски SNR - маркираните елементи в Таблица 2.2).

Подобно поведение се наблюдава и при ъгъла за GDS $\tau_g(\Omega)$ (Фиг. 2.11 (b) и (d)), само че в обратна посока - увеличението на β_g под влияние на шума е по-малко от това при β_c. Както и при редукцията на нормата, подобно изменение е по-силно изразено при SNR 0 dB.

Тези експериментално установени свойства на GDS вероятно са една от причините за получената по-висока точност при детекция на говор [Padmanabhan et al., 2008]. По-малкото изменение при зашумени сигнали на нормата на векторите и ъгъла между тях при GDS в сравнение с ЛП-кепстъра води до по-малко изменение под влияние на шума на съответните разстояния базиращи се на норми на вектори.

Представения тук анализ на спектъра на групово закъснение чрез адитивния спектрален модел е част от публикация на автора, отпечатана в сп. Автоматика и информатика [Узунов, 1993].

2.2.4. Group Delay Mean-Delta признак

GDS разгледан в предходната точка може да бъде дефиниран съгласно [Hegde et al., 2007] по следния аналогичен на формула (2.28) начин: Ако $x(n)$ е дискретизирания говорен сигнал то GDS $\tau(\omega)$ се представя във вида

$$\tau(\omega) = -\operatorname{Im}\frac{d(\log(X(\omega))}{d\omega} = \frac{X_R(\omega)Y_R(\omega)+Y_I(\omega)X_I(\omega)}{\left|X(\omega)\right|^2}, \qquad (2.43)$$

където със субскрипта R и I са означени реалната и имагинерната част на съответните преобразувания на Фурие. $X(\omega)$ and $Y(\omega)$ са преобразуванията на Фурие съответно за последователностите $x(n)$ и $nx(n)$. Съгласно резултатите представени в [Hegde et al.,

2007] е видно че в сравнение с амплитудния FFT спектър GDS се отличава с по-силно изразени спектрални пикове. Ако обаче в знаменателя в (2.43) е налице много малка стойност (spectrum's dip) то GDS се получава с прекалени силно изявени пикове, което съществено затруднява по нататъшното му приложение. Тези стойности в знаменателя за резултат от наличието на нули разположени близо до единичната окръжност в предавателната функция на гласовия тракт. За да се преодолее този недостатък при изчисляване на GDS са предложени различни подходи. При един от тях се изчислява т.н. Модифициран GDS (МСГЗ) или (Modified Group Delay Spectrum (MGDS)) и е предложен в [Hegde et al., 2007]. МСГЗ $\tau_m(\omega)$ е дефиниран като

$$\tau_m(\omega) = \left(\frac{\tau(\omega)}{|\tau(\omega)|} \right) (|\tau(\omega)|)^{\alpha} \tag{2.44}$$

където

$$\tau(\omega) = \left(\frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S(\omega)^{2\gamma}} \right) \tag{2.45}$$

и $S(\omega)$ е кепстрално изгладената версия на FFT спектъра $|X(\omega)|$. Параметрите α и γ се изменят 0 до 1 ($0 < \alpha \leq 1$) и ($0 < \gamma \leq 1$). Тези два параметъра и кепстрално изгладения спектър в знаменателя са въведени, за да редуцират амплитудните пикове и да ограничат динамичния диапазон на MGDS. За да се контролира степента на кепстрално изглаждане при $S(\omega)$ е използван кепстрален лифтер с дължина l_w.

В работата е предложен нов признак наречен Group Delay Mean-Delta (GDMD), който е предназначен за детекция на говор чрез контурен анализ. Той използва Mean-Delta подхода, предложен в т. 2.1.4, но в случая спектралната автокорелационна функция се определя не посредством FFT спектъра, а чрез модифицирания GDS дефиниран в (2.44). Основната цел на тази комбинация е да се използват свойствата на MGDS и да се постигне допълнително усилване на съответните пикове в делта спектралната автокорелационна функция.

В действителност са предложени две модификации на GDMD-признака. Първата условно е означена като lin-GDMD, а втората като log-GDMD. Създаването на двете модификации е поради факта, че при експерименталните изследвания бе извършено сравняване на ефективността на GDMD с тази на различни параметри, част от които са дефинирани в линейна скала, а друга част - в логаритмична. По-долу в текста при

описание на алгоритъма разликите между двете модификации ще бъдат изрично посочени.

За $n^{-я}$ сегмент предложените GDMD-признака се изчислява на три етапа (за по-голяма яснота индекса *n* е пропуснат в някои от формулите):

А. Първи етап – изчисляване на MGDS съгласно [Hegde et al., 2007], както следва:

- нека $x(n)$ е говорен сигнал в текущ сегмент, $n=1,\ldots,N$ е броят на отчетите в сегмента;
- прилагане на FFT върху последователностите *x(n)* и *nx(n)* и получаване на съответните спектри *X(k)* и *Y(k)*;
- определяне на $|S(k)|$ – кепстрално изгладения спектър на $|X(k)|$ чрез използване на лифтър с дължина l_w;
- определяне на MGDS $\tau_m(k)$ съгласно

$$\tau_m(k) = [sign].\left|\frac{X_R(k)Y_R(k)+Y_I(k)X_I(k)}{S(k)^{2\gamma}}\right|^{\alpha} \tag{2.46}$$

- където *[sign]* се определя от знака на израза

$$\frac{X_R(k)Y_R(k)+Y_I(k)X_I(k)}{S(k)^{2\gamma}}, \tag{2.47}$$

Стойностите на параметрите α, γ и l_w се определят експериментално.

Б. Втори етап – изчисляване на MD признака, използвайки MGDS $\tau_m(k)$ (2.46) от предходния етап, както следва:

- изчисляване на средния вектор на MGDS – използват се всички сегменти в анализираното произнасяне;
- прилагане на нормализация на MGDS за текущия сегмент спрямо средния вектор на MGDS (определен по цялото произнасяне);
- изчисляване на неизместената оценка на спектралната автокорелационна функция $R_m(l)$ използвайки нормализирания MGDS $\tau_m(k)$ от предходния етап

$$R_m(l) = \frac{1}{K/2-|l|}\sum_{k=0}^{K/2-1-l}\tau_m(k)\tau_m(k+l);\ l\geq 0;\ l\leq L; \tag{2.48}$$

  където *K* е размерът на FFT, *L* е броят на измествания (correlation lags) и *L=K/4*.
- изчисляване на делта спектралната автокорелационна функция $\Delta R_m(l)$ съгласно (2.6) използвайки $R_m(l)$ и делта коефициент *Q=3*

$$\Delta R_m(l) = \frac{\sum_{q=1}^{Q} q.(R_m(l+q) - R_m(l-q))}{2\sum_{q=1}^{Q} q^2} \tag{2.49}$$

- изглаждане на контура във времето на делта спектралната автокорелационна функция (за всеки лаг) чрез Long-Term Spectral Envelope (LTSE) algorithm с параметър *J=3* [Ramirez et al., 2004]. Така получената изгладена версия $\Delta R_m(n,l)$ е означена като $\Delta R_m^S(n,l)$

$$\Delta R_m^S(n,l) = \max\left\{\Delta R_m(n+j,l)\right\}_{j=-J}^{j=+J}. \tag{2.50}$$

- изчисляване на GDMD признака $m_{gd}(n)$ чрез $\Delta R_m^S(n,l)$ съгласно

$$m_{gd}(n) = \left[\sum_{l=0}^{L}\left|\Delta R_m^S(n,l)\right|\right] \tag{2.51}$$

B1. Трети етап 1–определяне lin-GDMD контура и изглаждане:

- изчисляване на lin-GDMD признака $m_{gd-lin}(n)$ чрез $m_{gd}(n)$ съгласно

$$m_{gd-lin}(n) = \left[m_{gd}(n)\right]^{0.5} \tag{2.52}$$

- изглаждане на m_{gd-lin} контура чрез филтър с изместваща се средна стойност;

B2. Трети етап 2– определяне на log-GDMD контура и изглаждане:

- нормализация на $m_{gd}(n)$ контура от (2.51) и получаване на крайния контур $m_{gd}^*(n)$ съгласно

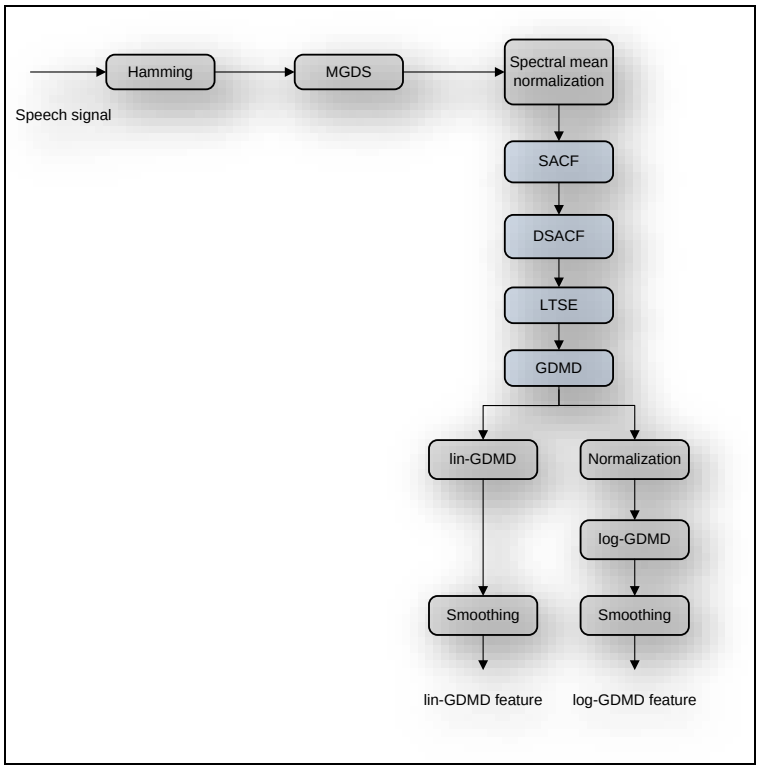
$$m_{gd}^*(n) = \left|m_{gd}(n) - m_{gd}^{\min}\right|, \tag{2.53}$$

където $m_{gd}^{\min} = \min_n\{m_{gd}(n)\}$.

- определяне на log-GDMD съгласно

$$m_{gd-log}(n) = log(1 + m_{gd}^*(n)) \tag{2.54}$$

- изглаждане на $m_{gd-\log}$ контура чрез филтър с изместваща се средна стойност.



Фиг. 2.12. Блок схема на алгоритъма за определяне на GDMD признаците.

На фиг. 2.12 е показана блок схемата на описания по-горе алгоритъм за определяне на GDMD признаците. Нормализацията, реализирана чрез ф-ли 2.53 и 2.54 е предложена поради обстоятелството, че получените минимални стойности в GDMD контура са винаги по-малки от 1, т.е. директното използване на логаритъм води до трудно интерпретируеми резултати.

2.3. Заключение

В първата част на Глава 2 са разгледани някои характеристики на спектралната автокорелационна функция, получена чрез FFT-спектъра. Предложен е метод, при който чрез прилагане на делта-филтър върху спектралната автокорелационна функция е получена т.н. делта спектрална автокорелационна функция. Демонстрирана е ефективността на тази филтрация, чрез която в спектъра на DSACF е постигнато значително усилване на спектралния пик, съответстващ на основния тон. Във втората част на главата са разгледани основните методи за определяне на GDS и е извършен качествен анализ на изменението на GDS при зашумени с адитивен шум говорни сигнали. Този анализ е реализиран косвено, чрез изследване на изменението на аргументите на проекционните функции на сходство на основата на адитивния спектрален модел. От една страна на базата само на свойствата на делта спектралната автокорелационна функция и от друга чрез комбинирането ѝ с модифицирания спектър

на групово закъснение са предложени общо пет признака за детекция на говор. Това са признаците – MD, log-GDMD, lin-GDMD, BMD и MMD. Първите три са предназначени за детекция чрез анализ на времеви контури, а последните два - за детекция чрез алгоритми за разпознаване.

2.4. Резюме на получените резултати към Глава 2

Научни резултати:

1. Предложен е метод, при който чрез прилагане на делта-филтър върху спектралната автокорелационна функция е получена т.н. делта спектрална автокорелационна функция. Анализирани са особеностите на тази филтрация, при която е постигнато усилване в честотната област на хармоничната структура на говорния сигнал. (Глава 2, т. 2.1.3).
2. Предложен е подход за изчисляване на признаци за детекция на говор базиращ се на свойствата на делта спектралната автокорелационна функция. Чрез този подход са дефинирани три признака. Първия от тях (т.н. MD-признак) е в скаларна форма и е предназначен за детекция чрез анализ на времеви контури, докато другите два (т.н. BMD- и MMD-признаци) са вектори и са предназначени за детекция чрез алгоритми за разпознаване (Глава 2, т. 2.1.4).
3. Извършен е теоретичен анализ на изменението на спектъра на групово закъснение при зашумени с адитивен шум говорни сигнали. Този анализ е реализиран косвено, чрез изследване на изменението на аргументите на проекционните функции на сходство на основата на адитивния спектрален модел (Глава 2, т. 2.2.3).
4. Предложен е подход за изчисляване на признаци за детекция на говор базиращ се комбинация на модифицирания спектър на групово закъснение и делта спектралната автокорелационна функция. Чрез този подход са дефинирани два признака (т.н. lin-GDMD и log-GDMD - признаци) които са предназначени за детекция на говор чрез анализ на времеви контури (Глава 2, т. 2.2.4).

ГЛАВА 3

Алгоритми за определяне на гранични точки при зависима от текста верификация на диктори. Експериментално изследване.

3.1. Увод

Една от основните причини за грешки в системите за автоматично разпознаване на говор и диктори е неправилното определяне на началната и крайната точки на анализираното говорно съобщение (дума или фраза) [Li et al., 2002].

В действителност модулът за Определяне на Граничните Точки (ОГТ) – Endpoint Detection (ED) е първи в една автоматична система за разпознаване. Грешно определените гранични точки водят в следващите етапи до ситуация, при която се обработва и разпознава или само част от говорното съобщение или към него се добавят непринадлежащи му фрагменти.

От една страна, ако работим с кратка говорна реализация, то елиминирането на част от нея поражда съществени различия между сравняваните модели и впоследствие увеличаване на грешката. От друга страна, ако към говорното съобщение се добавят излишни сегменти с шум и/или не говорни събития (и предполагайки, че те са част от него), то има голяма вероятност в генерирания модел съществено влияние да имат именно тези сегменти, а не истинския говор. Освен това добавянето на сегменти увеличава и времето за обработка [Li et al., 2002].

По-детайлно описание на алгоритмите за определяне граничните точки е представено в обзорната част на дисертацията поместена в т. 1.3.

В настоящата глава е извършен сравнителен експериментален анализ на ефективността на предложените в гл. 2 признаци предназначени за детекция на говор чрез анализ на времеви контури. Като референтни признаци са избрани: признак, получен чрез комбинация на енергията на сигнала и спектралната ентропия (Energy-Entropy (EE) parameter) [Huang et al., 2000]; спектрална ентропия с нормализиран спектър (Spectral Entropy with Normalized frame Spectrum – SENS parameter) [Renevey et al., 2001]; модифицирана енергия на Teager (Modified Teager's Energy – MTE parameter) [Gu et al., 2002] и дълговременна спектрална дивергенция (Long-term Spectral Divergence -LTSD parameter) [Luengo et al., 2010].

Необходимо е да се уточни, че *детектор, детектор на гранични точки* и *алгоритъм за определяне на гранични точки* са използвани в Глава 3 като синоними. Това е направено, за да се постигне по-голяма яснота на изложението.

В работата за даден признак е определено Евклидовото разстояние между неговите два Z-нормализирани времеви контура [Chen et al., 2005]. За конкретната фраза те са получени съответно от чистия ѝ запис и от зашумената ѝ версия. По този начин е налице количествена оценка на различията между контурите, предизвикани от влиянието на даден вид шум върху свойствата на конкретен признак. Използвани са записи на фрази на английски език избрани от корпуса SpEAR [SpEAR, online].

Точността на изследваните алгоритми за ОГТ е оценена чрез анализ на разликите между ръчно определените и получените от съответния алгоритъм гранични точки. Тестовете са реализиран с кратки фрази на български и английски език, които са избрани съответно от корпусите BG-SRDat [Ouzounov, 2003] и TIDIGITS [Dan Ellis, online].

Влиянието на алгоритмите за ОГТ върху точността на разпознаване е анализирано в две системи за зависима от текста верификация на диктори. При първата се използва алгоритъма Dynamic Time Warping (DTW) [Theodoridis et al., 2010], а при втората - скрити Марковски модели (Hidden Markov Models - HMMs) [Gales et al., 2008]. Тестовете са реализирани с кратки фрази на български език, записани по телефонен канал, и избрани от корпуса BG-SRDat [Ouzounov, 2003].

За да се оцени разликата (в статистически смисъл) между отделните алгоритми за детекция на граничните точки чрез използване на грешката при верификация е приложен Z_{HTER}-теста описан в [Bengio et al., 2004]. Изменението на грешката при верификация получена за всеки един от детекторите на гранични точки е визуализирано чрез усреднени DET графики [Beigi, 2011].

3.2. Референтни признаци

3.2.1. EE признак- Energy-Entropy (EE) feature

Известно е, че при детекция на говор в условията на нестационарен шум (например от механичен характер) спектралната ентропия е по-надежден признак от енергията на сигнала. Експериментите показват, обаче че при фонова музика и “babble noise” (съпътстващи разговори) нейните стойности се отличават със силна вариативност [Huang et al., 2000]. Този факт не позволява при детекция тя да се използва като самостоятелен признак. Имено, затова в [Huang et al., 2000] е предложен признак за ОГТ, който е комбинация между спектралната ентропия и енергията на говорния сигнал. Тази

комбинация е направена с цел да се преодолеят недостатъците на всеки един от тях и да се формира нов признак по-устойчив към някои от видовете шум.

За $n^{-я}$ сегмент ЕЕ-признака се определя по следния начин:

-определяне на енергията на сигнала $E(n)$

$$E(n) = \sum_{i=0}^{I-1} x(i)^2 , \tag{3.1}$$

където I е броя дискретни стойности в сегмента.

-определяне на функцията на плътността на вероятността $P(n,k)$ за честотната компонента k като

$$P(n,k) = \frac{|X(n,k)|^2}{\sum_{k=0}^{K/2} |X(n,k)|^2}, \tag{3.2}$$

където K е размерът на FFT.

- определяне на отрицателната спектрална ентропия

$$H(n) = \sum_{k=0}^{K/2} P(n,k).log\,2(P(n,k)) \tag{3.3}$$

- определяне на ЕЕ признака

$$EE(n) = \left[1 + |E(n) \times H(n)|\right]^{0.5} \tag{3.4}$$

3.2.2. Модифицирана енергия на Teager

Енергията на Teager е мярка за енергията на сигнала, която включва информация не само за амплитудата, но и за честотата на сигнала. Това нейно свойство я прави по-чувствителна (в сравнение с традиционната енергия на сигнала) при разграничаване на съгласни звуци с високочестотно съдържание [Li et al., 2012]. В непрекъсната времева област тя се дефинира чрез диференциални уравнения от втори ред, а в дискретната област за $i^{-тия}$ отчет има вида

$$E(i) = x^2(i) - x(i+1)\,x(i-1) \tag{3.5}$$

В [Gu et al., 2002] е предложено вместо дискретните стойности да се изчисляват съгласно формула (3.5) и след това да се определи енергията на сегмента, да се използва спектъра на мощността. В този случай параметъра модифицирана енергия на Teager - Modified Teager's Energy - MTE $E_t(n)$ за $n^{-я}$ сегмент има вида

$$E_t(n) = \left[\sum_{k=0}^{K/2} (k\Delta f)^2 |X(k)|^2\right]^{0.5}, \tag{3.6}$$

където Δf е разрешението по честота, $|X(k)|^2$ е FFT спектъра на мощността и K е размерът на FFT.

3.2.3. Спектрална ентропия с нормализиран спектър

Известно е, че контура във времето на спектралната ентропия в участъци от сигнала зашумени с цветен шум е много сходен с контура в участъци, в които няма говор [Renevey et al., 2001]. За да се реализира детекция на говор с помощта на спектралната ентропия в участъци, където е налице цветен шум е предложено в [Renevey et al., 2001] спектъра във всеки сегмент да се нормализира спрямо средния спектър определен по всички сегменти в изследваното произнасяне.

Ако $|X(n,k)|$ е амплитудният спектър за n-я сегмент, където $n = 0,...,N-1$; $k = 0,...,K/2$ и K е размерът на FFT и N е броят на сегментите в произнасянето, то нормализирания спектър $|\hat{X}(n,k)|$ е изчислен съгласно

$$|\hat{X}(n,k)| = \frac{|X(n,k)|}{\frac{1}{N}\sum_{n=0}^{N-1}|X(n,k)|}. \tag{3.7}$$

Функцията на плътността на вероятността $P(|\hat{X}(n,k)|^2)$ за спектъра $|\hat{X}(n,k)|$ се определя чрез нормализация на честотните компоненти

$$P(|\hat{X}(n,k)|^2) = \frac{|\hat{X}(n,k)|^2}{\sum_{k=0}^{K/2}|\hat{X}(n,k)|^2}, \tag{3.8}$$

и спектралната ентропия с нормализиран спектър (Spectral Entropy with Normalized frame Spectrum (SENS)) $H_w(n)$ за n-я сегмент е

$$H_w(n) = -\sum_{k=0}^{K/2} P(|\hat{X}(n,k)|^2).\log(P(|\hat{X}(n,k)|^2)). \tag{3.9}$$

Отрицателната SENS $H_w^-(n)$ е дефинирана като $H_w^-(n) = -H_w(n)$.

3.2.4. Дълговременна спектрална дивергенция

LTSD (Long-Term Spectral Divergence) параметъра е предложен в [Ramirez at al., 2004] и е дефиниран като отклонение на дълговременната спектрална обвивка спрямо средния спектър на шума.

Ако $x(n,i)$ е дискретизиран говорен сигнал, където $n = 0,...N-1$, N е броят сегменти и $i = 0,...,I-1, I$ е броят на дискретните стойности в сегмента и амплитудния

спектър $|X(n,k)|$ на *x(n,i)* е получен чрез FFT където $k = 0,...,K-1$ и *K* е размера FFT. Дълговременната спектрална обвивка (Long-Term Spectral Envelope - LTSE) от *M*-ти ред е дефинирана като

$$LTSE_M(n,k) = max\left\{|X(n+j,k)|\right\}_{j=-M}^{j=+M} \tag{3.10}$$

Дълговременната спектрална дивергенция (Long-Term Spectral Divergence - LTSD) от *M*-ти ред се определя като отклонение на $LTSE_M$ по отношение на средния амплитуден спектър на шума $|S(k)|$ или

$$LTSD_M(n) = 10\log_{10}\left(\frac{1}{K}\sum_{k=0}^{K-1}\frac{LTSE_M^2(n,k)}{|S(k)|^2}\right), \tag{3.11}$$

Алгоритъма е адаптивен спрямо променящото се ниво на шума чрез обновяване (updating) на амплитудния спектър на шума извършващо се само за неговорни сегменти. Известни са различни версии на LTSD алгоритъма, но тук ще бъде използвана версията, описана в [Luengo et al., 2010, §2]. В алгоритъма е дефинирана калибрираща функция, определена чрез анализ на енергията на шума при чисти и при зашумени фрагменти локализирани по цялата база данни използвана в изследванията. Ако праговете γ_0 и γ_1 съответстват на енергийните нива E_0 и E_1 които са средните стойности на енергията съответно за чист и зашумен сигнал, то прагът γ се адаптира за всеки сегмент съгласно

$$\gamma(n) = \begin{cases} \gamma_0 & E_N(n) \le E_0 \\ \dfrac{\gamma_0-\gamma_1}{E_0-E_1}E_N(n) + \gamma_0 - \dfrac{\gamma_0-\gamma_1}{1-\dfrac{E_1}{E_0}} & E_0 < E_N(n) < E_1 \\ \gamma_1 & E_N(n) \ge E_1 \end{cases} \tag{3.12}$$

където $E_N(n)$ е енергията на шума за *n-я* сегмент и се дефинира като

$$E_N(n) = \begin{cases} \alpha E_N(n-1) + (1-\alpha)E(n); & if\ non\ speech \\ E_N(n-1); & if\ speech \end{cases}, \tag{3.13}$$

където $E(n)$ е енергията на сигнала за *n-я* сегмент.

Амплитудния спектър на шума $|S(n,k)|$ се адаптира съгласно

$$|S(n,k)| = \begin{cases} \alpha|S(n-1,k)| + (1-\alpha)|X(n,k)|; & if\ non\ speech \\ |S(n-1,k)|; & if\ speech \end{cases} \tag{3.14}$$

Началните стойности $E_N(0)$ в (3.13) и $|S(0,k)|$ в (3.14) са средната енергия и средния спектър на шума и са изчислени от началните сегменти на съответния файл, предполагайки, че там е налице само шум.

3.3. Анализ на Z-нормализирани времеви контури

Тук е прието като груба мяра (rough measure) на влиянието на шума върху времевия контур за даден признак да се използва големината на сходство между съответните контури на чистия говор и неговата зашумена версия. Сходството във формата (shape similarity) може да бъде определено чрез Евклидовото разстояние между двата нормализирани контура [Chen et al., 2005]. Прието е да се използва Z-нормализация, при която нормализирания контур има средна стойност 0 и средно квадратично отклонение 1. Този тип нормализация е възможно да се приложи, защото контурите са с еднаква дължина и липсват локално-времеви изместванията. Освен това чрез така полученото разстояние е налице количествена оценка на структурните различия на контурите предизвикани от влиянието на даден вид шум върху свойствата на конкретен признак.

Нормализирания контур $T_Z(n)$ където $n=1,...N$ и N е броят на сегментите се определя съгласно

$$T_Z(n)=\frac{T(n)-m_T}{\sigma_T}, \tag{3.15}$$

където m_T, σ_T са съответно средната стойност и средно квадратичното отклонение, изчислени за целия контур $T(n)$.

Експериментите в тази подточка са разделени в две групи. В първата група се включват предложени в гл. 2 и референти признаци, при които в дефинирането им не се използва логаритмична функция – условно ще бъдат назовани линейни признаци. Във втората група са тези с логаритмична функция. Разделянето в две групи е по технически причини, а именно поради трудностите с едновременната визуализация на голям брой графики. В първата група са включени следните признаци:

- базов MD признак (Basic MD feature) – т.2.1.4.1 и фиг.2.10 (предложен);
- линеен GDMD признак (lin-GDMD feature) – т.2.2.4 и 2.12 (предложен);
- Energy-Entropy (EE) признак – т.3.2.1, (референтен);
- Modified Teager's Energy – MTE – т.3.2.2, (референтен);
- Spectral Entropy with Normalized frame Spectrum – SENS – т.3.2.3, (референтен).

Във втората група са включени признаци, при които в дефинирането им е използвана допълнителна нормализация и логаритмична функция. Използването на логаритъм е с цел да се усилят участъци от контура с ниски стойности, които са характерни за слаби фрикативни съгласни. Тези признаци са:

- логаритмичен базов MD признак (log-Basic MD feature) – получава се като вместо корен квадратен във формула 2.17 се приложат формули 2.53 и 2.54 по аналогичен начин на този използван при определяне на log-GDMD контура в т.2.2.4 (предложен);
- логаритмичен GDMD признак (log-GDMD feature) – т.2.2.4 и фиг. 2.12 (предложен);
- логаритмичен Energy-Entropy (EE) признак – т.3.2.1, във ф-ла 3.4 вместо корен квадратен се използва логаритъм (референтен);
- логаритмичен Modified Teager's Energy – MTE – т.3.2.2; във ф-ла 3.6 вместо корен квадратен се използва логаритъм (референтен);
- LTSD признак – т.3.2.4 (референтен).

Логаритмична версия на признака SENS не се използва при тестовете поради демонстрираната висока вариативност на базовата версия, дефинирана в т. 3.2.3 (виж таблица 3.1).

Избрани са следните записи от секцията "Lombard Speech" на корпуса SpEAR:

- заводски шум – включва говор съдържащ заводски шум и записан в заводско хале. Чистият запис е със SNR=27.28 dB, а зашумената версия е със SNR=-9.96 dB.

  Текст – *"Eric has an automobile factory in this house"*.

  Чист запис: *t_alex_factoryR1_8.wav*

  Зашумен запис: *alex_factoryR1_8.wav*
- шум от движещ се автомобил – включва говор съдържащ шум от движеща се кола. Чистият запис е със SNR=27 dB, а зашумената версия е със SNR=-14.58 dB.

  Текст – *"We are going to have Easter brunch"*.

  Чист запис: *t_michele_mvolvoR1_8.wav*

  Зашумен запис: *michele_mvolvoR1_8.wav*
- розов шум - включва говор, записан при акустично генериран розов шум. Чистият запис е със SNR=21.23 dB, а зашумената версия е със SNR=-10.33 dB.

  Текст – *"I'm sitting in a room with pink noise in the green rag"*.

  Чист запис: *t_alex_pinkR5_8.wav*

  Зашумен запис: *alex_pinkR5_8.wav*

- шум от кабината на F16 - включва говор съдържащ шум от кабината на посочения самолет. Чистият запис е със SNR=24.40 dB, а зашумената версия е със SNR=-1.05 dB.

  Текст – *"This is an example of Lombard speech for SpEAR database".*

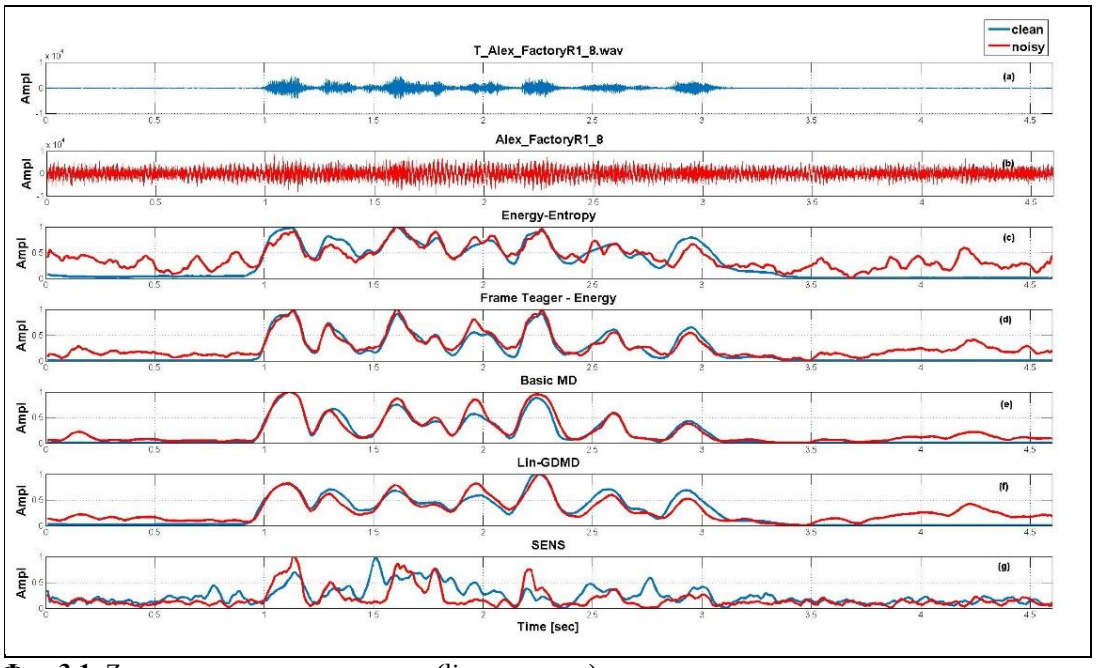
  Чист запис: *t_alex_f16noiseR2_8.wav*

  Зашумен запис: *alex_f16noiseR2_8.wav*

Всички записи са направени в реална среда (live speech in a noisy environment) [SpEAR, online]. Оригиналните записи от корпуса SpEAR са с честота на дискретизация 16 kHz. При използваните в изследването говорни данни тази честота е редуцирана на 8 kHz.

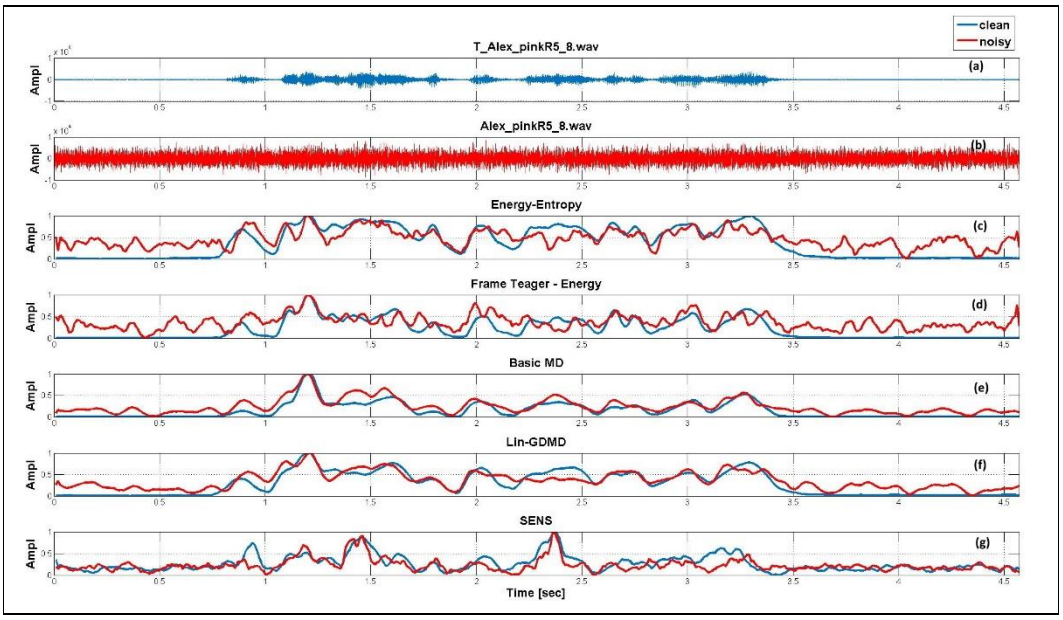
3.3.1. Експерименти с линейни признаци

На фиг. 3.1 са показани Z-нормализираните контури на чистия и зашумения сигнал за запис със заводски шум при SNR= - 9.96 dB и петте избрани признака, съответно – (a) чист сигнал; (b) зашумен сигнал; (c) Energy-Entropy (d) Modified frame Teager-Energy; (e) Basic MD; (f) lin-GDMD и (g) SENS.



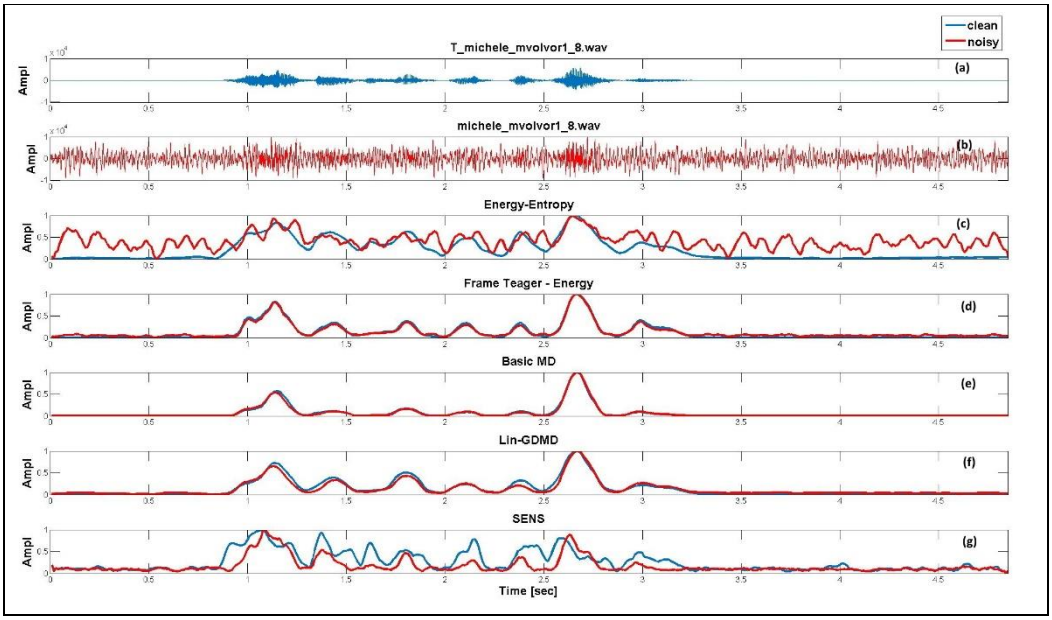
Фиг.3.1. Z-нормализираните контури (lin-признаци) на чистия и зашумения сигнал за запис със заводски шум.

На фиг. 3.2 са показани Z-нормализираните контури на чистия и зашумения сигнал за запис с розов шум при SNR = -10.33 dB и петте избрани признака, съответно – (a) чист сигнал; (b) зашумен сигнал; (c) Energy-Entropy (d) Modified frame Teager-Energy; (e) Basic MD; (f) lin-GDMD и (g) SENS.



Фиг.3.2. Z-нормализираните контури (lin-признаци) на чистия и зашумения сигнал за запис с розов шум.

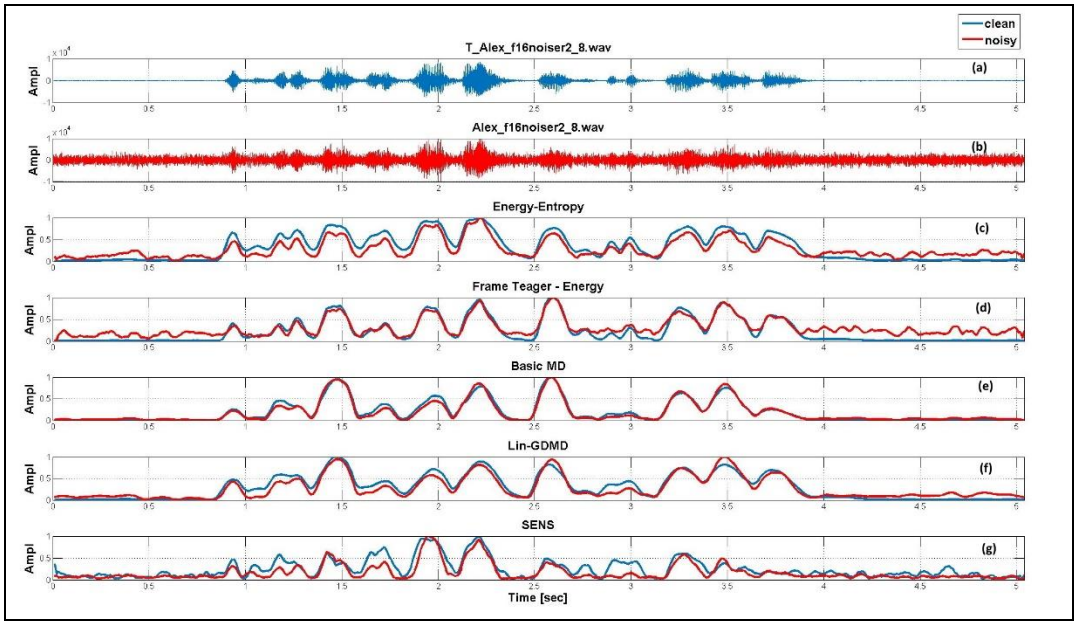
На фиг. 3.3 са показани Z-нормализираните контури на чистия и зашумения сигнал за запис с шум от движещ се автомобил при SNR = -14.58 dB и петте избрани признака, съответно – (a) чист сигнал; (b) зашумен сигнал; (c) Energy-Entropy (d) Modified frame Teager-Energy; (e) Basic MD; (f) lin-GDMD и (g) SENS.



Фиг.3.3. Z-нормализираните контури (lin-признаци) на чистия и зашумения сигнал за запис с шум от движещ се автомобил.

На фиг. 3.4 са показани Z-нормализираните контури на чистия и зашумения сигнал за запис с шум от кабината на F16 при SNR = -1.05 dB и петте избрани признака,

съответно – (a) чист сигнал; (b) зашумен сигнал; (c) Energy-Entropy (d) Modified frame Teager-Energy; (e) Basic MD; (f) lin-GDMD и (g) SENS.



Фиг. 3.4. Z-нормализираните контури (lin-признаци) на чистия и зашумения сигнал за запис на шум от кабината на F16.

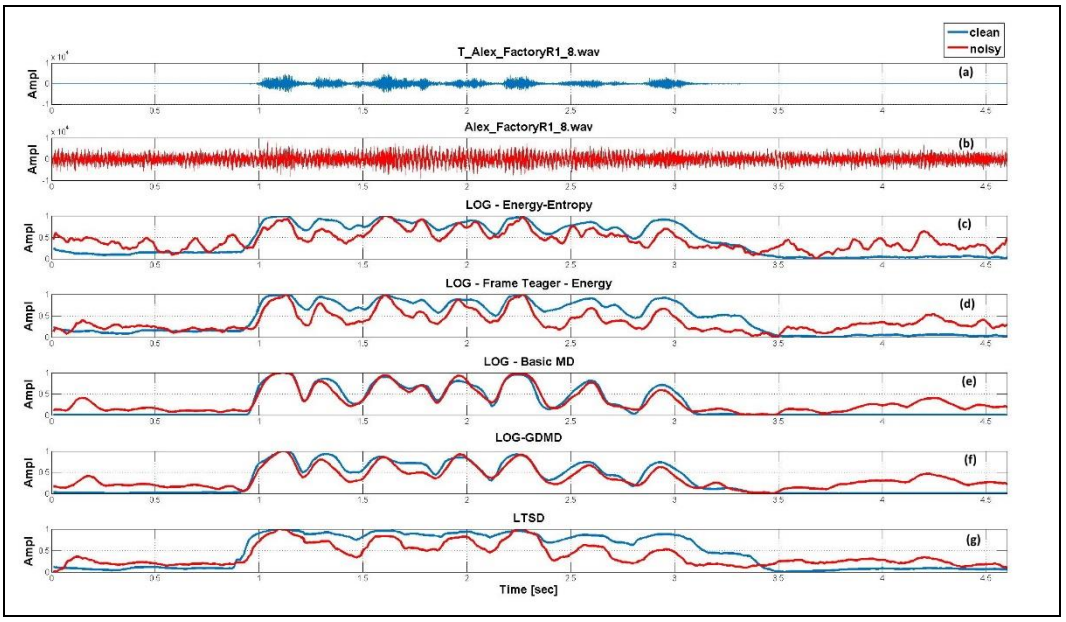
В таблица 3.1 за всеки един признак са показани средните Евклидови разстояния между нормализираните контури на записите избрани от корпуса SpEAR. Минималната стойност на разстоянието за всеки запис е визуализирана с удебелен шрифт.

Таблица 3.1. Средни Евклидови разстояния между Z-нормализираните контури (lin-признаци)

		Noise			
No.	**Parameters**	**Pink**	**Car**	**Factory**	**F16**
1	**EE**	0.0327	0.0384	0.0263	0.0182
2	**MTE**	0.0392	0.0071	0.0184	0.0153
3	**Basic MD**	**0.0205**	**0.0044**	**0.0130**	**0.0080**
4	**Lin-GDMD**	0.0241	0.0098	0.0223	0.0149
5	**SENS**	0.0353	0.0327	0.0429	0.0213

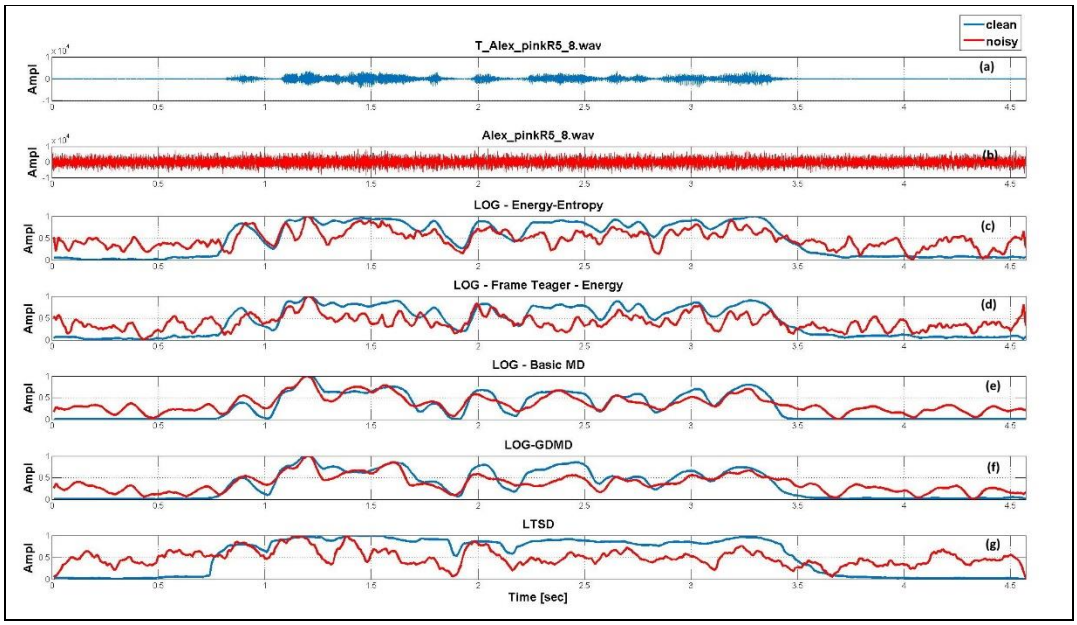
3.3.2. Експерименти с логаритмични признаци

На фиг. 3.5 са показани Z-нормализираните контури на чистия и зашумения сигнал за запис със заводски шум при SNR=-9.96 dB и петте избрани признака, съответно – (a) чист сигнал; (b) зашумен сигнал; (c) log-Energy-Entropy (d) log-Modified frame Teager-Energy; (e) log-Basic MD; (f) log-GDMD и (g) LTSD.



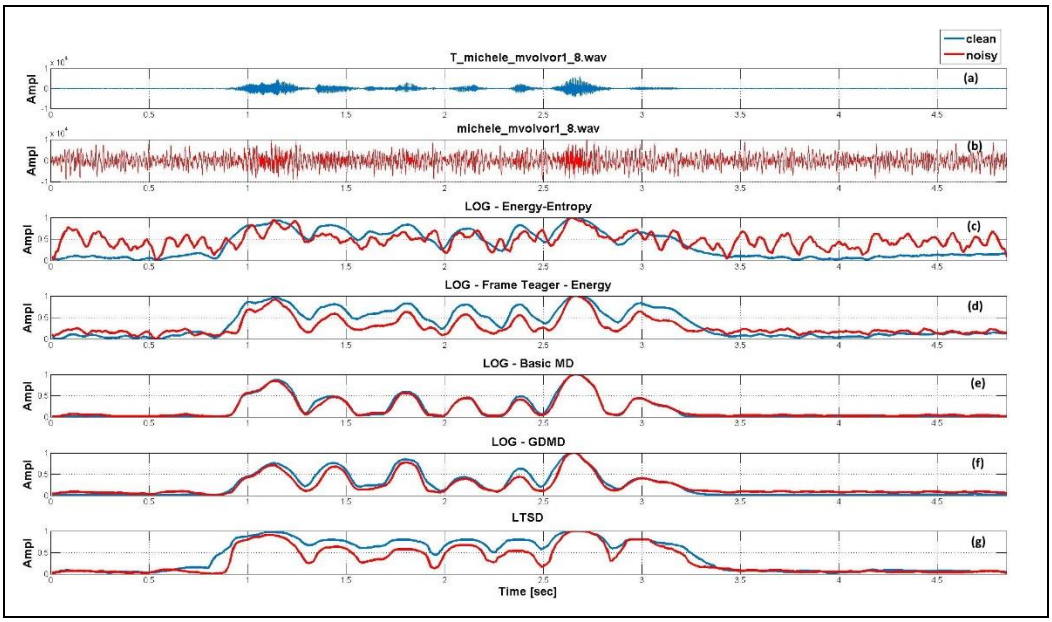
Фиг. 3.5. Z-нормализираните контури (log-признаци) на чистия и зашумения сигнал за запис със заводски шум.

На фиг. 3.6 са показани Z-нормализираните контури на чистия и зашумения сигнал за запис с розов шум при SNR=-10.33 dB и петте избрани признака, съответно – (a) чист сигнал; (b) зашумен сигнал; (c) log-Energy-Entropy (d) log-Modified frame Teager-Energy; (e) log-Basic MD; (f) log-GDMD и (g) LTSD.



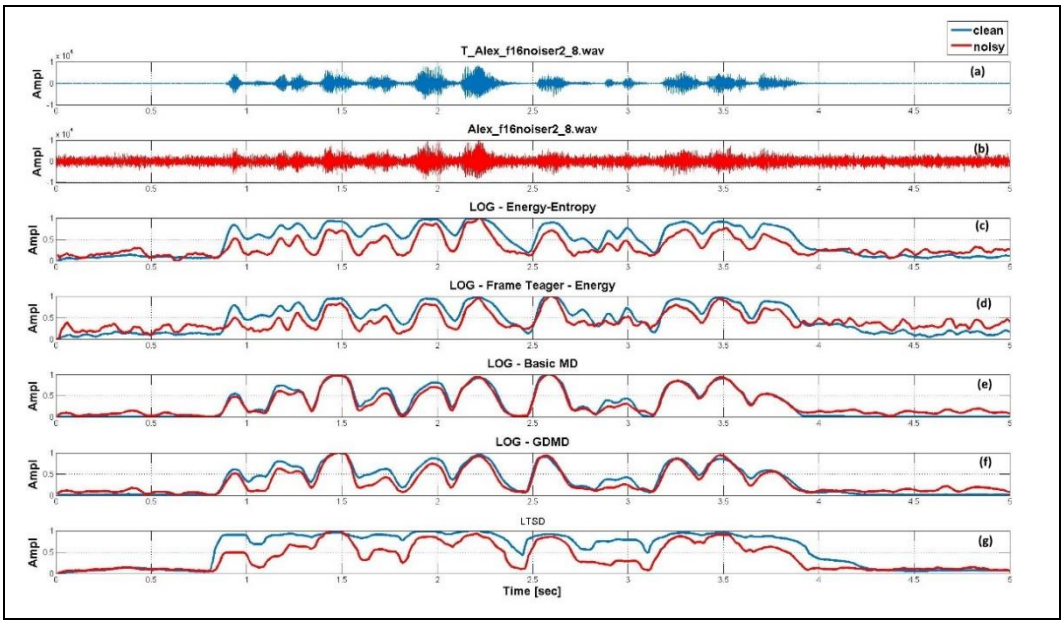
Фиг. 3.6. Z-нормализираните контури (log-признаци) на чистия и зашумения сигнал за запис с розов шум.

На фиг. 3.7 са показани Z-нормализираните контури на чистия и зашумения сигнал за запис с шум от движещ се автомобил при SNR=-14.58 dB и петте избрани признака, съответно – (a) чист сигнал; (b) зашумен сигнал; (c) log-Energy-Entropy (d) log-Modified frame Teager-Energy; (e) log-Basic MD; (f) log-GDMD и (g) LTSD.



Фиг. 3.7. Z-нормализираните контури (log-признаци) на чистия и зашумения сигнал за запис с шум от движещ се автомобил.

На фиг. 3.8 са показани Z-нормализираните контури на чистия и зашумения сигнал за запис с шум от кабината на изтребител F16 при SNR=-1.05 dB и петте избрани признака, съответно – (a) чист сигнал; (b) зашумен сигнал; (c) log-Energy-Entropy (d) log-Modified frame Teager-Energy; (e) log-Basic MD;(f) log-GDMD и (g) LTSD.



Фиг. 3.8. Z-нормализираните контури (log-признаци) на чистия и зашумения сигнал за запис на шум от кабината на F16.

В таблица 3.2 са всеки един log-параметър са показани средните Евклидови разстояния между нормализираните контури на записите избрани от корпуса SpEAR.

Минималната стойност на разстоянието за всеки запис е визуализирана с удебелен шрифт.

Таблица 3.2. Средни Евклидови разстояния между Z-нормализираните контури (log-признаци)

		Noise			
No.	**Parameters**	**Pink**	**Car**	**Factory**	**F16**
1	**Log-EE**	0.0344	0.0412	0.0307	0.0257
2	**Log-MTE**	0.0421	0.0232	0.0339	0.0274
3	**Log-Basic MD**	**0.0237**	**0.0063**	**0.0173**	**0.0102**
4	**Log-GDMD**	0.0269	0.0128	0.0235	0.0160
5	**LTSD**	0.0511	0.0161	0.0284	0.0238

3.3.3. Дискусия и заключение

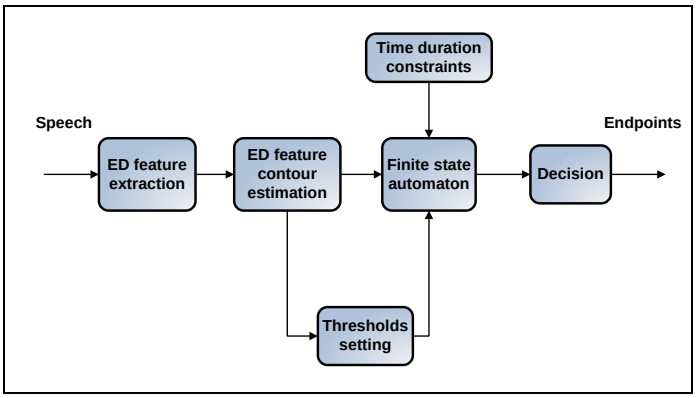
Получените експериментални резултати демонстрират ясно предимствата на предложените признаци – Basic MD и GDMD (и в двете им версии – линейна и логаритмична). В сравнение с всички останали признаци при Basic MD Евклидово разстояние има винаги минимална стойност, т.е. налице са минимални различия между контурите на този признак за чистия и зашумения сигнал. Този факт улеснява работата на алгоритмите за определяне на праговите стойности, които са описани по-долу в текста. При съществени различия между контурите е много трудно чрез прагови стойности да се локализират говорни фрагменти - например подобен случай е показан на фиг. 3.3 (c). Необходимо е да се уточни, че получените резултати имат индикативен (указателен) характер и са сериозен мотив при по-нататъшното използване на предложените MD-признаци в алгоритмите за детекция на говор.

3.4. Алгоритми за определяне на граничните точки

Най-често при разработка на детектори за определяне на гранични точки на кратки фрази се реализира съвместна работа на два алгоритъма - първия за изчисляване на прагови стойности (фиксирани или адаптивни) и втория за управление на прагово-времевите съотношения в анализирания контур чрез краен автомат [Li et al., 2002], [Abdulla et al., 2009], [Chung et al., 2014]. В работата е предложен подход за разработване на такъв детектор включващ алгоритъм за изчисляване на адаптивни прагове и детерминиран краен автомат. Първоначалната версия на предложения краен автомат е разработена от автора като част от алгоритъм за определяне на гранични точки, описан в [SR-API, 2003]. Тук е представена версия, която е частично адаптирана към целите и задачите на текущите изследвания. В повечето случаи подобни детектори за ОГТ са *ad hoc* решения.

В процеса на изследователската работа бе установено, че е изключително трудно точно да се възпроизведат решения, които се базират на евристични правила. Поради това в дисертацията ефективността на предложения краен автомат е сравнена с hangover алгоритъма, който е добре описан в стандарта [ETSI, 2007].

На базата на предложения подход и в зависимост от характеристиките на контурите на признаците са разработени три алгоритъма за определяне на гранични точки. Общата блок схемата на подобен алгоритъм за ОГТ е показан на фиг. 3.9.



Фиг. 3.9. Блок схема на алгоритъм за определяне на гранични точки.

3.4.1. Алгоритъм за определяне на фиксирани прагове

Известни са два начина при определяне на фиксирани прагови стойности. При първия се приема, че в началото на анализирания аудио сигнал съществува фрагмент, в който не е наличен говор. Този фрагмент (нарича се още период на адаптация) се използва за определяне на параметрите на наличния шум и в последствие те служат за изчисляване на праговите стойности. Проблеми при подобен начин за определяне на праговете възникват, ако през периода на адаптация се появи говорен сигнал или шума е нестационарен и е с високо ниво в този участък. Това води до изчисляване на неадекватни прагови стойности и впоследствие до груби грешки при определяне на граничните точки [Zhang et al., 2010]. При втория начин не се правят предположения къде са разположени говорните фрагменти, а се анализира целия аудио запис.

В работата е предложен алгоритъм за определяне на две фиксирани прагови стойности T_{low} и T_{high} - сходен алгоритъм е описан в [Kitaoka et al., 2007] където се използва само един фиксиран праг и базов праг, определен чрез алгоритъма на Otsu. Предложения алгоритъм има вида:

Step 1. Изчисляване на стойностите на избрания параметър $C(n), n = 1, \cdots N$ където N е броят на сегментите в анализираното произнасяне;

Step 2. Определяне на базовия праг T_{base} като средната стойност на $C(n)$;

$$T_{base} = \mathbf{E}\{C(n)\}; \tag{3.16}$$

Step 3. Изчисляване на средната стойност m_{down} като

$$m_{down} = \mathbf{E}\{C(n) < T_{base}, n = 1, \cdots N\}; \tag{3.17}$$

Step 4. Изчисляване на средната стойност m_{up} като

$$m_{up} = \mathbf{E}\{C(n) \geq T_{base}, n = 1, \cdots N\}; \tag{3.18}$$

Step 5. Проверка - Ако m_{down} има стойност близка до нула, то

$$\text{if } \frac{m_{down}}{m_{up}} < \gamma \text{ then } m_{down} = \gamma m_{up}; \tag{3.19}$$

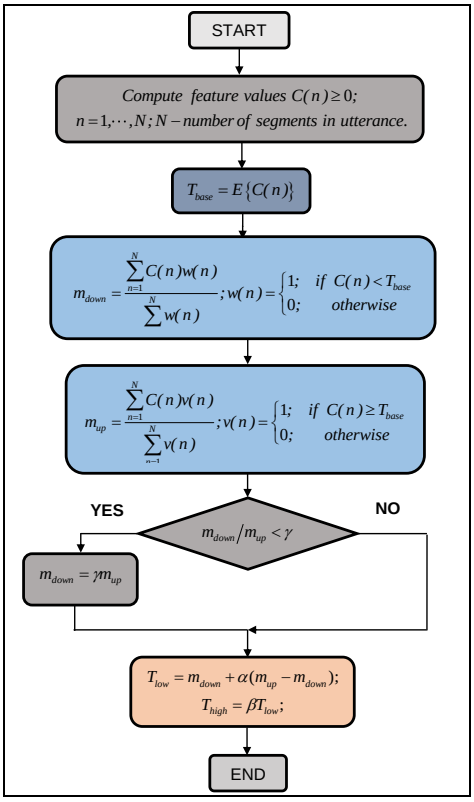
Step 6. Изчисляване на ниския (първия) праг T_{low}

$$T_{low} = m_{down} + \alpha(m_{up} - m_{down}); \tag{3.20}$$

Step 7. Изчисляване на високия (втория) праг T_{high}

$$T_{high} = \beta T_{low}; \tag{3.21}$$

Коефициентите α, β и γ се определят експериментално (виж т. 3.6.2). Блок схемата на описания алгоритъм е показана на фиг. 3.10.



Фиг. 3.10. Блок схема на алгоритъма за определяне на фиксираните прагове.

3.4.2. Алгоритъм за определяне на адаптивни прагове

В началото на произнасянето преходът от беззвучен към звучен фрагмент предизвиква рязко увеличение в стойностите на контура (например ако използвания параметър има характеристики сходни с тези на енергията на сигнала). От друга страна в края на произнасянето преходът от звучен към беззвучен фрагмент се характеризира с бавно намаление в стойностите на контура. Много често подобни беззвучни фрагменти грешно се класифицират като шум. Така описаното изменение в контура затруднява съществено използването на фиксирани прагови стойности в алгоритмите за детекция.

За да се намали грешката при детекция на говор чрез използване на фиксирани прагове тук е предложен алгоритъм - подобрена версия в посока адаптивност на вече описания в т. 3.4.1 - при който се изчисляват две двойки от прагови стойности. Първата двойка е предназначена за определяне на началната гранична точка на произнасянето, а втората – за крайната точка. Или казано по друг начин чрез анализ на контурните характеристики в началото на произнасянето се определят два фиксирани прага, и тези прагове се използват от крайния автомат само за определяне на началната гранична точка. Допълнително, чрез анализ на контурните характеристики в края на произнасянето, се определя втора двойка от прагове, която се прилага само за определяне на крайната гранична точка. Ключовия проблем при предложения алгоритъм е как да се локализират началния и крайния участъци от произнасянето, като се използват само характеристиките на контура. Тук се предлага решението да бъде намерено чрез анализ на пиковите стойности в контура.

Ако $P=\{p_i\}, i=1,\ldots,G,$ е множество от пикови стойности, където G е общият брой на пиковете в анализирания контур. Всеки пик е дефиниран като $p_i=(a_i,l_i)$ където a_i е амплитудата на пика и l_i е неговото местоположение, т.е. номера на сегмента, където той е разположен. Дефинира се ново множество $Q_M=\underset{a}{\text{sort}}\{P\}$ получено след сортиране на амплитудите на пиковете a_i в низходящ ред и се изберат първите M и $M<<G$. Определят се $l_{\min}$ и $l_{\max}$ където $l_{min}=\underset{l}{min}\{Q_M\}$ и $l_{max}=\underset{l}{max}\{Q_M\}$. Местоположението на разделящата точка l_{spl}, т.е., точката (номер на сегмент) която разделя контура на две части – начална и крайна – е дефинирано като $l_{\text{spl}}=l_{\min}+\kappa(l_{\max}-l_{\min})$.

В предложения алгоритъм за всяка част на контура е изчислен т.н. начален праг. Чрез този праг се определят две средни стойности m_{down} и m_{up}. Първата средна стойност

е изчислена чрез стойностите на контура по-малки от началния праг, а втората – от стойностите по-големи или равни на същия праг. Използвайки тази идея са определени две двойки прагове, за началната част на контура $T_{\text{beg}}^{\text{low}}, T_{\text{beg}}^{\text{high}}$ и за крайната му част - $T_{\text{end}}^{\text{low}}, T_{\text{end}}^{\text{high}}$. Предложения алгоритъм има вида:

Step 1. Изчисляване на стойностите на контура $C(n) \geq 0; n = 1, \ldots N$, N е броят на сегментите;

Step 2. Намиране на пиковете в контура $P = \{ p_i \}; p_i = (a_i, l_i);$ a_i е амплитудата на пика, l_i е локацията на пика и $i = 1, \ldots, G$, G е броят на пиковете.

Step 3. Получаване на последователност от подредените в низходящ ред пикове $Q_M = \underset{a}{\text{sort}}\{P\}$ и избор на първите M от тях; $M << G$;

Step 4. Определяне на $l_{\min} = \min\limits_{l}\{Q_M\}$ и $l_{\max} = \max\limits_{l}\{Q_M\}$;

Step 5. Изчисляване на точката (сегмента) на разделяне:

$$l_{\text{spl}} = l_{\min} + \kappa(l_{\max} - l_{\min}). \tag{3.22}$$

Step 6. Изчисляване на началните прагове за началната и крайната част на контура:

$$\begin{aligned} T_{\text{beg}}^{\text{init}} &= \text{E}\{C(n)\}; \;\; n = 1, \ldots, l_{\text{spl}}, \\ T_{\text{end}}^{\text{init}} &= \text{E}\{\text{C}(n)\}; \;\; n = l_{\text{spl}} + 1, \ldots, N. \end{aligned} \tag{3.23}$$

Step 7. Изчисляване на допълнителните средни стойности за началната част на контура:

$$m_{\text{beg}}^{\text{down}} = \frac{\sum\limits_{n=1}^{l_{\text{spl}}} C(n)w(n)}{\sum\limits_{n=1}^{l_{\text{spl}}} w(n)}, \;\; w(n) = \begin{cases} 1 & \text{if } C(n) < T_{\text{beg}}^{\text{init}}, \\ 0 & \text{otherwise,} \end{cases} \tag{3.24}$$

$$m_{\text{beg}}^{\text{up}} = \frac{\sum\limits_{n=1}^{l_{\text{spl}}} C(n)v(n)}{\sum\limits_{n=1}^{l_{\text{spl}}} v(n)}, \;\; v(n) = \begin{cases} 1 & \text{if } C(n) \geq T_{\text{beg}}^{\text{init}}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.25}$$

Step 8. Изчисляване на допълнителните средни стойности за крайната част на контура:

$$m_{\text{end}}^{\text{down}} = \frac{\sum\limits_{n=l_{\text{spl}}+1}^{N} C(n)w(n)}{\sum\limits_{n=l_{\text{spl}}+1}^{N} w(n)}, \;\; w(n) = \begin{cases} 1 & \text{if } C(n) < T_{\text{end}}^{\text{init}}, \\ 0 & \text{otherwise,} \end{cases} \tag{3.26}$$

$$m_{\text{end}}^{\text{up}} = \frac{\sum_{n=l_{\text{spl}}+1}^{N} C(n)v(n)}{\sum_{n=l_{\text{spl}}+1}^{N} v(n)}, \quad v(n) = \begin{cases} 1 & \text{if } C(n) \geq T_{\text{end}}^{\text{init}}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.27}$$

Step 9. Изчисляване на двойката прагови стойности за началната част на контура:

$$\begin{aligned} T_{\text{beg}}^{\text{low}} &= m_{\text{beg}}^{\text{down}} + \alpha_1 (m_{\text{beg}}^{\text{up}} - m_{\text{beg}}^{\text{down}}), \\ T_{\text{beg}}^{\text{high}} &= \max(T_{\text{beg}}^{\text{init}}, \beta_1 T_{\text{beg}}^{\text{low}}). \end{aligned} \tag{3.28}$$

Step 10. Изчисляване на двойката прагови стойности за крайната част на контура:

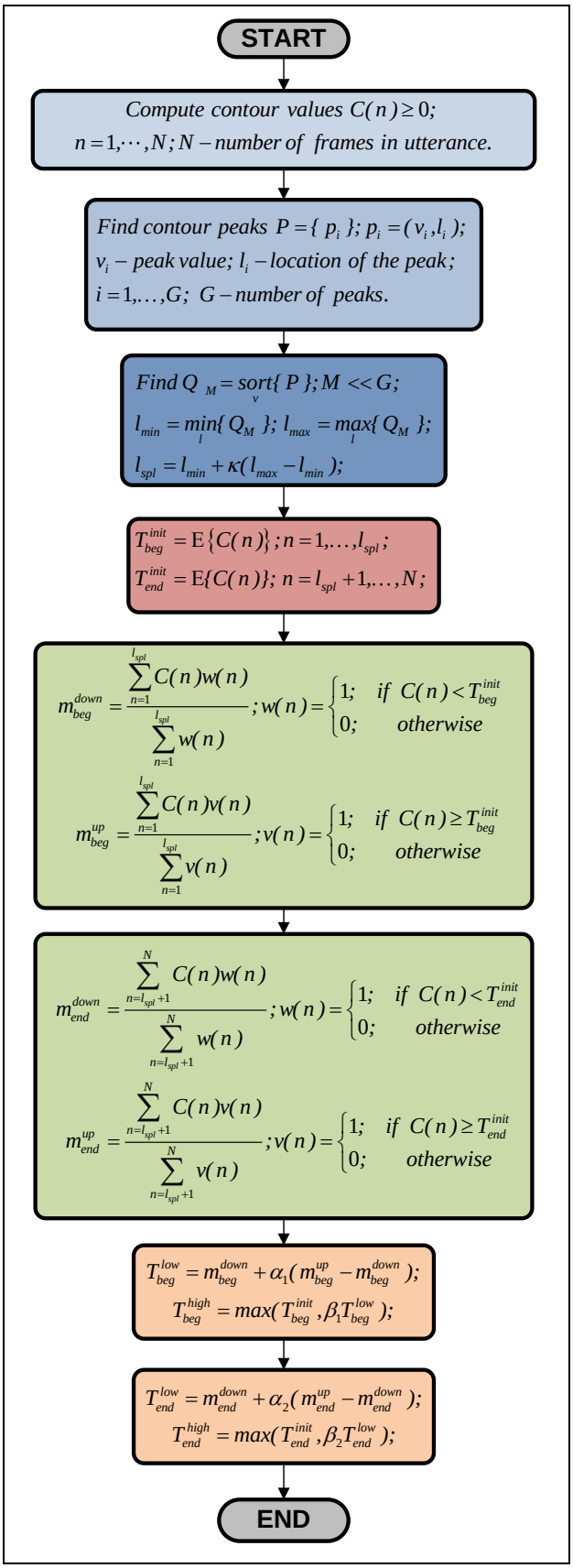
$$\begin{aligned} T_{\text{end}}^{\text{low}} &= m_{\text{end}}^{\text{down}} + \alpha_2 (m_{\text{end}}^{\text{up}} - m_{\text{end}}^{\text{down}}), \\ T_{\text{end}}^{\text{high}} &= \max(T_{\text{end}}^{\text{init}}, \beta_2 T_{\text{end}}^{\text{low}}). \end{aligned} \tag{3.29}$$

Параметрите $\alpha_1, \beta_1, \alpha_2, \beta_2, \kappa$ и M подлежат на настройка съобразно условията на експеримента (виж т. 3.6.2). Блок схемата на описания алгоритъм е показана на фиг. 3.11.

3.4.3. Детерминиран краен автомат. Описание.

В [Тилков и Бояджиев, 1977] са представени обстойни изследвания в областта на българската фонетика и в този източник се твърди, че в българския език няма думи, които да започват с повече от четири съгласни и също така няма думи, които да завършват с повече от три съгласни. Реализираните предварителни изследвания с множество от думи, избрани от [Тилков и Бояджиев, 1977], показаха следното: звучните фрагменти могат да бъдат предшествани (в началото на думата) и следвани (в края на думата) от беззвучни такива с дължина от порядъка на 200-400 и 400-600 ms. Необходимо е да се уточни, че за английския език се твърди, че няма думи, които да започват с повече от три съгласни и също така няма думи, които да завършват с повече от четири съгласни [Roach, 2009]. Освен това описаните по горе фрагменти в началото и в края на думата са с продължителност за английския език съответно 300 и 500 ms [Ghaemmaghami et al., 2010b]. Извършването на подробен анализ на този проблем обаче не е цел на настоящето изследване.

Посочените по-горе две времеви константи се използват в разработения от автора детерминиран краен автомат за дефиниране дължините на областите преди звучните фрагменти в началото на думата и след звучните такива в края на думата, където ще се извършва търсенето на гранични точки. Тъй като тези константи за българския и английския езици са близки по стойност то само една двойка константи ще бъде използвана в изследванията.



Фиг. 3.11. Блок схема на алгоритъма за определяне на адаптивните прагове.

Предложения краен автомат е с 8 състояния, съответно: INIT, SCAN_DATA, SCAN_START, MAYBE_IN, SCAN_END, MAYBE_OUT, END_FOUND и END. Отличителна характеристика на автомата е, че при определени условия той генерира съобщение за грешка. Ако това се случи, алгоритъма за детекция прекратява работата си и текущите аудио данни не постъпват към следващите нива на системата за разпознаване. Грешка се генерира при следните условия:

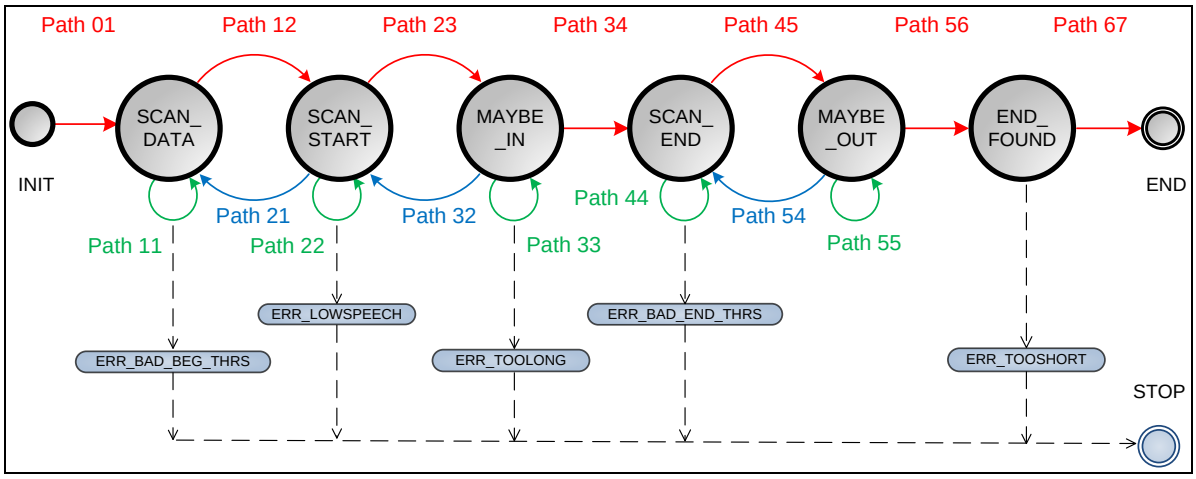
- когато произнасянето завършва извън аудио файла (не се изпълняват условията за край на говорното съобщение) - error ERR_TOOLONG;
- когато SNR е много ниско - error ERR_LOWSPEECH;
- когато праговете не позволяват да бъдат определени началните или крайните точки (лошо установени прагове) - errors ERR_BAD_BEG_THRS, ERR_BAD_END_THRS;
- когато дължината на произнасянето е по-малка от предварително дефинираната минимална продължителност (за избягване на звукови артефакти, генерирани от диктора) - error ERR_TOOSHORT.

Този механизъм за контрол на грешката е въведен, за да не се допуска неподходящи говорни реализации да бъдат въведени в системата за разпознаване. Защитата от т.н. неподходящи произнасяния или звукови артефакти е важен етап в системите за верификация на диктори работещи в реално време и по телефонен канал.

Логиката на предложения краен автомат е показана на фиг. 3.12. Съответните правила за преход са изброени в таблица 3.3. Параметрите $T_{\text{SCAN\_START}}$, $T_{\text{MAYBE\_IN}}$, $T1_{\text{SCAN\_END}}$, $T2_{\text{SCAN\_END}}$ и $T_{\text{MAYBE\_OUT}}$ са таймери показващи колко време автоматът е в дадено състояние (state timers). Въведени са времеви константи *MaxQuietTime, UpTime1, UpTime2, MiddleTime, MinLengthTime, EndTime, BegTime, MaxStateTime*, които от една страна служат за ограничаване на минималната и максималната продължителност на произнасянето, на дължината на вътрешните паузи, а от друга определят времевите условия за преминаване на автомата в следващо състояние.

В предложения автомат са въведени два вида т.н. *Endpoint Candidates (EC)* – това са номера на сегменти, които са потенциални кандидати за гранични точки. ЕС са съответно type-0 и type-1. Вида на ЕС зависи от няколко фактора. Например type-0 в края на произнасянето е типично за беззвучни фрагменти. В този случай времевия период между последния ЕС type-1 и последния type-0 не може да бъде по-голям от времевата константа *EndTime*. В таблица 3.3 флагът EoF се установява, когато е достигнат края на файла (End-of-File).

Като илюстрация на резултатите от предложения алгоритъм на фиг. 3.13 е показан контура на признака log-GDMD за зашумен сигнал от корпуса SpEAR (същият сигнал е показан на фиг. 3.7). Времевата диаграма на преходите е показан на фиг. 3.13 (c). По дължината на контура във фиг. 3.13 (d) са означени по-важни детайли в работата на алгоритъма: ръчно и автоматично определените гранични точки; точката на разделяне при определяне на адаптивните прагове; *EC type-1* и двойките прагове за определяне респективно на началните $T_{\text{beg}}^{\text{low}}$, $T_{\text{beg}}^{\text{high}}$ и крайните точки $T_{\text{end}}^{\text{low}}$, $T_{\text{end}}^{\text{high}}$. Необходимо е да се уточни, че тук е описана версията на крайния автомат, която използва алгоритъма с адаптивни прагове в т. 3.3.2.



Фиг. 3.12. Диаграма на логиката на крайния автомат.

В текста под $C(n)$ се разбира стойността в $n^{-я}$ сегмент на признака, чийто контур ще бъде анализиран. Състоянията и условията за преход на крайния автомат са както следва:

- ***INIT*** – *Установяване на стойностите по подразбиране за всички параметри;* Установяване като работни прагове T_{low} и T_{high} на двойката прагове за определяне на началната точка T_{beg}^{low} и T_{beg}^{high}.
- ***SCAN_DATA*** – *Търсене на кандидати за начална гранична точка;* Анализира се контура и се търси сегмента, при който $C(n) \geq T_{low}$. Ако има такъв, номера на сегмента се маркира като кандидат за начална точка и се преминава към следващото състояние. Ако няма и е достигнат края на аудио файла се генерира грешка– *ERR_BAD_BEG_THRS*.
- ***SCAN_START*** – *Анализ на контура между двете прагови стойности;*

Ако $T_{low} \le C(n) < T_{high}$ за период по-дълъг от *MaxQuietTime*, се генерира грешка – *ERR_LOWSPEECH*. Ако $C(n) < T_{low}$ алгоритъма се връща в предходното състояние. Ако $C(n) \ge T_{high}$ се преминава към следващото състояние.

Таблица 3.3. Правила за преход на крайния автомат

Paths	State Transition	Rules of State Transition	Errors
01	**INIT→ SCAN_DATA**	Set the beginning pair as work thresholds T_{low} and T_{high}. Set parameters' default values and go to SCAN_DATA.	
11	**SCAN_DATA → SCAN_DATA**	Stay in SCAN_DATA until ($C(n)< T_{low}$).	If current state is SCAN_DATA and EoF is set, then ERR_BAD_BEG_THRS
12	**SCAN_DATA→ SCAN_START**	Go to SCAN_START if ($C(n)\ge T_{low}$) – ***n*** is marked as beginning point candidate.	
21	**SCAN_START→ SCAN_DATA**	Go back to SCAN_DATA if ($C(n)< T_{low}$).	
22	**SCAN_START→ SCAN_START**	Stay in SCAN_START until ($C(n)\ge T_{low}$) & ($C(n)< T_{high}$) & ($T_{SCAN\_START} \le$ *MaxQuietTime*).	If ($C(n)\ge T_{low}$) & ($C(n)< T_{high}$) & ($T_{SCAN\_START} >$ *MaxQuietTime*) then ERR_LOWSPEECH
23	**SCAN_START→ MAYBE_IN**	Go to MAYBE_IN if ($C(n)\ge T_{high}$).	
32	**MAYBE_IN → SCAN_START**	Go back to SCAN_START if ($T_{MAYBE\_IN} <$ *UpTime2*) & ($C(n)< T_{high}$).	
33	**MAYBE_IN → MAYBE_IN**	Stay in MAYBE_IN until ($T_{MAYBE\_IN} <$ *UpTime2*) & ($C(n)\ge T_{high}$).	If current state is MAYBE_IN and EoF is set, then ERR_TOOLONG
34	**MAYBE_IN → SCAN_END**	if ($T_{MAYBE\_IN} \ge$ *UpTime2*) - estimate the beginning point *BPoint* using *BegTime* – then go to SCAN_END.	
44	**SCAN_END → SCAN_END**	Stay in SCAN_END until ($C(n)\ge T_{low}$). If **($n \ge l_{spl}$)** then set the ending pair as work thresholds T_{low} and T_{high} .	If current state is SCAN_END, EoF is set and no EC, then ERR_BAD_END_THRS
45	**SCAN_END → MAYBE_OUT**	Go to MAYBE_OUT if ($C(n)< T_{low}$) – ***n*** is marked as ending point candidate.	
54	**MAYBE_OUT→ SCAN_END**	Go back to SCAN_END if (($C(n) \ge T_{high}$) & ($T1_{SCAN_END} \ge$UpTime1)) OR (($C(n) \ge T_{low}$) & ($T2_{SCAN_END} \ge$MiddleTime)).	
55	**MAYBE_OUT → MAYBE_OUT**	Stay in MAYBE_OUT until (($C(n) \ge T_{high}$) & ($T1_{SCAN\_END} <$ *UpTime1*)) OR (($C(n) \ge T_{low}$) & ($T2_{SCAN\_END} <$ *MiddleTime*)).	
56	**MAYBE_OUT→ END_FOUND**	if ($C(n)< T_{low}$) & ($T_{MAYBE\_OUT} \ge$ *MaxStateTime*) - estimate ending point *EPoint* and utterance length *ULength* - then go to END_FOUND	
67	**END_FOUND→ END**	Go to END if (*ULength* $\ge$ *MinLengthTime*).	If (*ULength* $<$ *MinLengthTime*) then ERR_TOOSHORT

- ***MAYBE_IN*** – *Определяне на началната гранична точка;*

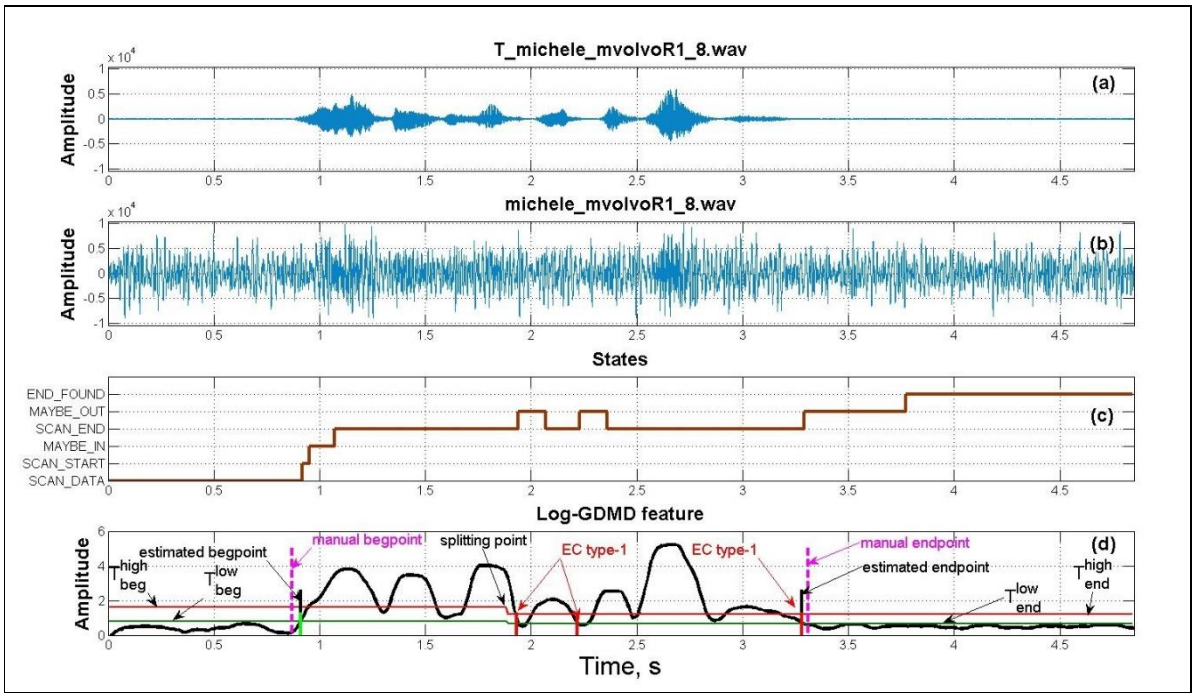
  Ако $C(n) < T_{high}$ за период по-малък от *UpTime2* автомата се завръща в предходното състояние. Ако този период е равен или по-голям от *UpTime2*, то се счита, че е налице говорно съобщение и се маркира съответния сегмент, в който $C(n) \geq T_{high}$. В този случай се анализират кандидатите за начални точки разположени в период с продължителност *MaxBegTime* преди въведения маркер и се определя действителната начална точка *BPoint*. След което се преминава към следващото състояние. Ако е достигнат края на аудио файла и текущото състояние все още е MAYBE_IN, то се генерира грешка – *ERR_TOOLONG*.
- ***SCAN_END*** – *Търсене на крайна гранична точка;*

  Ако номера на текущия сегмент е равен или по-голям от номера на сегмента определен като точка на разделяне ($n \geq l_{spl}$ - ф-ла 3.22), то двойката прагове за крайната точка T_{end}^{low} и T_{end}^{high} се установяват като работни прагове T_{low} и T_{high}. Търсят се сегменти, при които $C(n) < T_{low}$. Ако има, то номера на сегмента се маркира като първи кандидат за крайна точка и се преминава към следващото състояние. Ако е достигнат края на аудио файла и не е намерен кандидат за крайна точка, то се генерира грешка – *ERR_BAD_END_THRS*.
- ***MAYBE_OUT*** – *Определяне на края на произнасянето;*

  Анализират се стойности на контура разположени между двата прага за период по-голям от *MiddleTime* или стойности по-големи или равни на T_{high} за период равен или по-голям от *UpTime1*. Ако има такава стойност, то тя се маркира като следващ кандидат за крайна точка и алгоритъма се връща в предходното състояние. Ако няма за период по-голям от *MaxStateTime*, то се преминава към следващото състояние.
- ***END_FOUND*** – *Проверка на дължината на произнасянето;*

  Извършва се анализ на кандидатите за крайна гранична точка и се определя действителната крайна точка – *EPoint*. Изчислява се дължината на произнасянето *ULength=EPoint-BPoint*. Ако *ULength* ≥ *MinLengthTime* то се преминава към следващото състояние. Ако не, то се генерира грешка – *ERR_TOOSHORT*.
- ***END*** – *Край;*

Алгоритъма завършва успешно и изпраща определените гранични точки към следващия етап на обработка в системата за разпознаване.

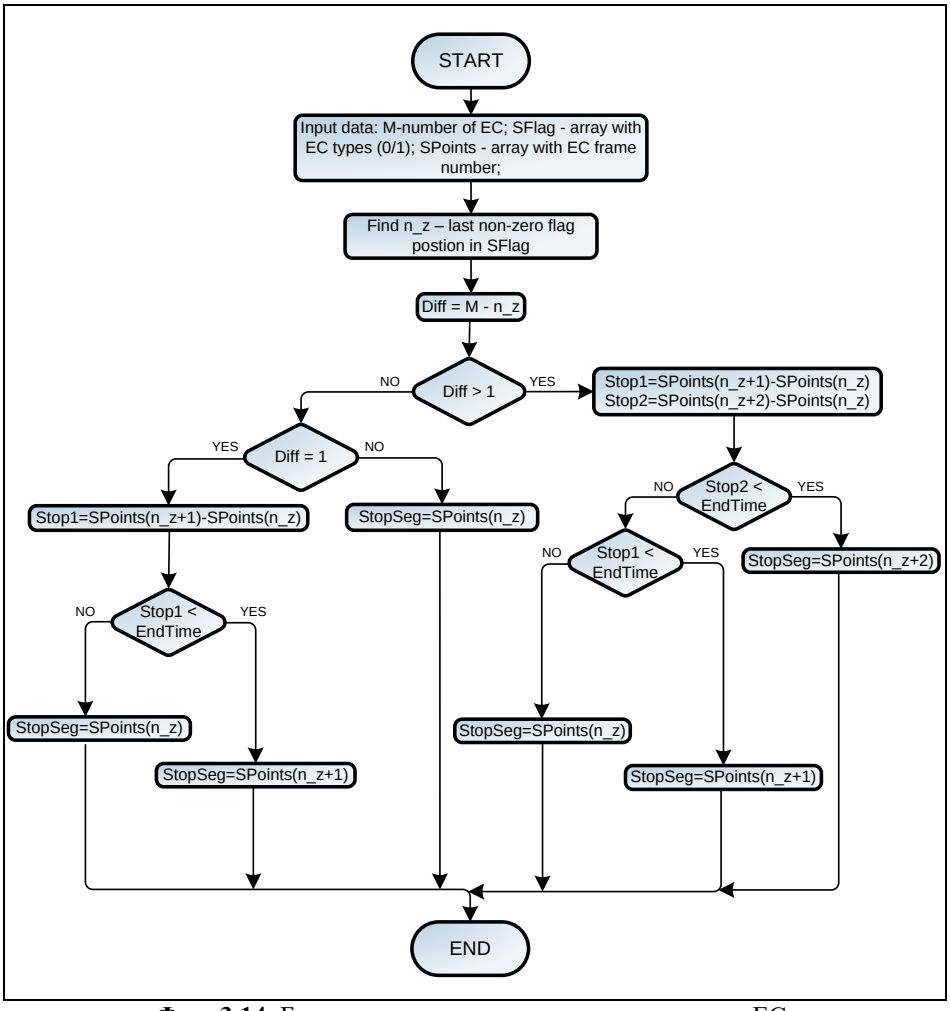


Фиг. 3.13. Запис от корпуса SpEAR: (a) чист сигнал; (b) зашумена версия; (c) времева диаграма на преходите; (d) контур на log-GDMD признака с маркирани някои характерни детайли от изпълнението на предложения алгоритъм във времето.

Алгоритмите за детекция на граничните точки грешат най-често при определяне на края на произнасянето. Причината е специфичното изменение на контура в края на думата при преход от звучни към беззвучни фрагменти, както и наличието на звукови артефакти, породени от диктора. В предложения алгоритъм е въведен анализ на кандидатите за крайни гранични точки, в който се вземат предвид тези особености.

В алгоритъма за въведени два вида (Endpoint Candidates) EC – type-0 и type-1. Като EC е обозначен сегмент (при състояние на крайния автомат ≥ SCAN_END), в който стойността на контура става по-малка от T_{low}. Типа на EC се определя от това какъв вид е предходното пресичане на праговете T_{low} и T_{high} от страна на контура. Ако предходното пресичане на контура е на праг $T_{high,}$ то имаме EC type-1. Другия тип EC type-0 се получава ако предходното пресичане е на праг $T_{low.}$ Както бе посочено по-горе EC type-0 е характерен за съгласни разположени в края на думата, когато е възможно думата да завърши с многократни колебания на контура между двата прага, без да се надвишава втория праг T_{high}. В предложения алгоритъм се анализират два EC след последния EC type-1. Ако няма такива то последния EC type-1 е крайната гранична точка. Ако има и те са [type-0; type-0], то се определя времевия период между последния EC type-1 и всеки един EC от двойката. Тези периоди са означени съответно като Stop1 и Stop2. Ако Stop2 < *EndTime*, то втория EC type-0 се приема за крайна точка. Ако не, проверява се дали Stop1< *EndTime*. Ако да, първия EC type-0 се приема за крайна точка. Ако не е изпълнено

времевото ограничение и за първия EC type-0, то като крайна точка се приема последния ненулев EC. По подобен начин се процедира и ако след последния EC type-1 имаме само един EC type-0. Константата *EndTime* е 500 ms, и съгласно ограниченията описани в началото на т. 3.4.3 това е максималният времеви период след звучни фрагменти в края на произнасянето, където може да се търси крайна гранична точка. Блок схемата на описания алгоритъм за анализ на EC е показана на фиг. 3.14.



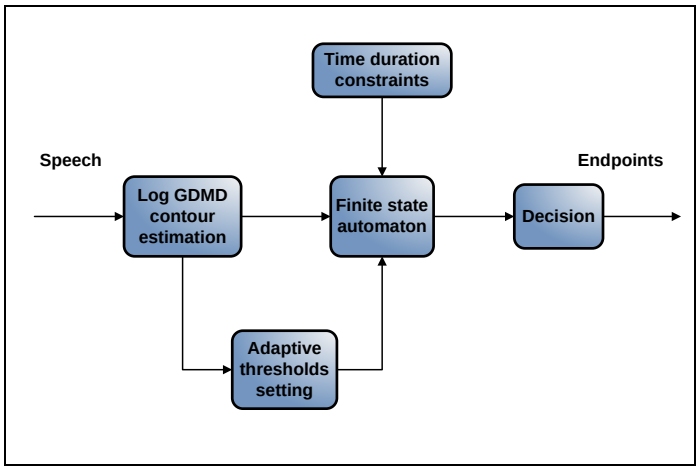
Фиг. 3.14. Блок схемата на алгоритъма за анализ на EC.

3.5. Детектори на гранични точки

3.5.1. Детектор GDMD-E

Този детектор е комбинация от признака log-GDMD (т. 2.2.4), алгоритъма за установяване на адаптивни прагове (т. 3.3.2) и предложения краен автомат (т. 3.3.3). Първо се изчислява времевия контур на log-GDMD признака. След което контура се анализира и се определят двете двойки с прагови стойности. Използвайки така

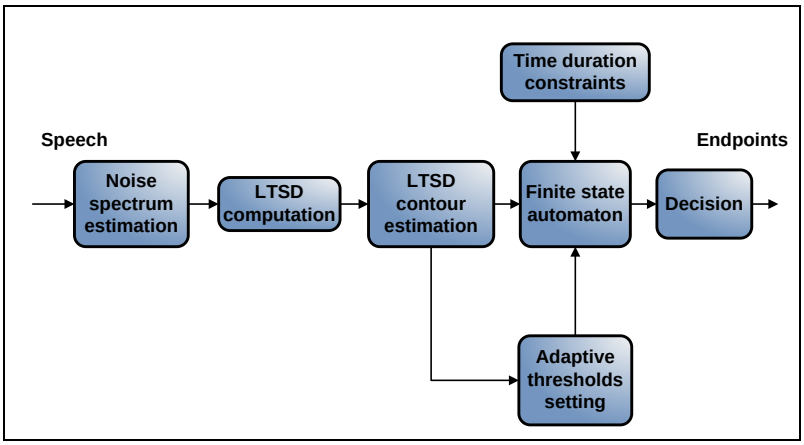
определените прагове и логиката на крайния автомат се намират граничните точки на анализираното произнасяне. Блок схемата на GDMD-E детектора е показана на фиг. 3.15.



Фиг. 3.15. Блок схема на детектор GDMD-E.

3.5.2. Детектор LTSD-E

Този тип детектор е предложен, за да се изследва ефективността на признака LTSD. Обикновено при него се прилага единичен праг и hangover схема [ETSI, 2007]. Тук вместо обичания подход се използват адаптивни прагове и краен автомат аналогично на детектора от предходната точка. Блок схемата на LTSD-E детектора е показана на фиг. 3.16.

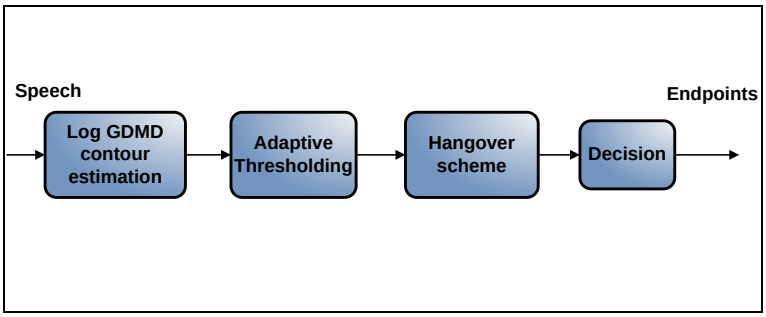


Фиг. 3.16. Блок схема детектор LTSD-E.

3.5.3. Детектор GDMD-H

Този детектор е предложен, за да се анализира ефективността на hangover алгоритъма при използването му с контура на log-GDMD. Първоначално се определя контура и се изчисляват двете двойки прагове. Hangover схемата обикновено се използва за

изглаждане на последователността 0/1 (не говор/говор) получена от модул за детекция на говор. За да се получи подобна последователност е решено да се приложи алгоритъма с адаптивните прагове като само високите прагове се използват (като единичен праг) за генериране на последователността 0/1. В действителност чрез генериране на бинарна последователност се реализира класическия VAD алгоритъм, но тук само първия и последния сегменти с говор се локализират и те се приемат като гранични точки. Блок схемата на предложения детектор GDMD-H е показан на фиг. 3.17.



Фиг. 3.17. Блок схема на детектор GDMD-H.

3.6. Експерименти

В работата са реализирани две групи от експерименти, чиято цел е да се оцени ефективността на предложените алгоритми за детекция на гранични точки. При първата група е изследвана точността на детекция чрез анализ на разликите между ръчно определените гранични точки и тези получени от предложените алгоритми. При втората група е установено как различните алгоритми за детекция влияят върху точността на разпознаване при зависима от текста верификация на диктори.

3.6.1. Говорни данни

Говорните данни, използвани в експериментите са избрани от два корпуса - BG-SRDat [Ouzounov, 2003] и TIDIGITS [DanEllis, online]. При първата група от експерименти – оценка на точността - са използвани данни и от двата корпуса, докато при втората група –верификация на диктори - данни само от първия корпус.

Корпуса BG-SRDat съдържа фрази на български език записани от зашумени телефони канали. Данните са дискретизирани с честота 8 kHz при 16 бита, PCM-формат, и mono режим. Дължината на единичен запис (файл) е около 2.5-3 сек., като дължината на самата фраза е около 2 сек. Данните, използвани в това изследване включват 337 файла и са събрани от 18 диктора (мъже). Записите са осъществени при реални условия (real-world environment) което води до ситуация, при която някои от записите са

сравнително чисти, докато други са изключително зашумени. Необходимо е да се доуточнят някои детайли за използваната фраза. Фразата е: "*Здравей Манолов! Как се чувстваш днес?"*. Тя е предложена от автора в [Ouzounov, 2003] (виж Глава 5 на дисертацията) и характерно за нея е, че започва с две шумови звучни съгласни *'зд'* и завършва с шумова беззвучна съгласна *'с'*. Този факт затруднява съществено определянето на граничните точки при телефонни записи.

От корпуса TIDIGITS, който е на английски език, са избрани записи съдържащи последователности от цифри. Всички записи са формирани като двойка сигнали - чист сигнал и съответни зашумени версии, получени чрез филтрация, реверберация и добавяне на шум с различни нива. Данните са дискретизирани с честота 8 kHz при 16 бита, PCM-формат, mono режим и включват 84 файла, събрани от 6 диктора (3 мъже и 3 жени).

За всички записи, използвани в експериментите е извършено ръчно определяне на граничните точки на фразите.

3.6.2. Параметри на алгоритмите. Настройки.

Параметрите в предложените алгоритми, които подлежат на настройка се установяват само в експериментите за определяне на точността при детекция. Техните стойности се избират така, че разпределението на разликата (в брой сегменти) между ръчно и автоматично определени гранични точки да има максимум в диапазона под 10 сегмента. Така получените параметри в последствие се използват при разпознаване на диктори.

Стойностите на основните параметри са: тегловна функция на Хеминг – дължина 30 ms, изместване 10 ms, точки на FFT- 512. При изчисляване на log-GDMD - α= 0.6; γ = 0.4; l_w = 32 в ф-ли (2.46) и Q=3; J=6 в ф-ли (2.49) и (2.50).

По време на експериментите се появиха проблеми със стойностите на някои параметри в LTSD-алгоритъма. В него се прилага калибрираща функция, за да се установи адаптивен праг. Използването на стойностите на параметрите на калибриращата функция предложени в [Luengo et al., 2010], [Ramirez et al., 2004] доведе до лоши резултати при всички тестове. Подобни наблюдения за тези стойности са направени и в [Tuononen et al., 2008]. Този факт наложи извършването на интензивни експериментални изследвания за да бъдат намерени нови стойности на съответните параметри за LTSD –алгоритъма. Тези стойности са:

- TIDIGITS: $E_0 = 70; E_1 = 90; \gamma_0 = 15; \gamma_1 = 10; \alpha = 0.95; \Delta_{\text{offset}} = 0;$
- BG-SRDat: $E_0 = 60; E_1 = 90; \gamma_0 = 20; \gamma_1 = 6; \alpha = 0.95; \Delta_{\text{offset}} = 2.$

Началната оценка на средния спектър на шума е изчислена от спектрите на първите 10 сегмента от файла, предполагайки, че там има само шум.

При hangover алгоритъма се използват евристични правила включващи таймери и прагови стойности по дължина (в брой сегменти), които са същите както предложените в [ETSI, 2007] или B=7; S_P=3; S_L=4; L_S=5 и L_M=23, където B е дължина на буфер, S_P е праг, при който е възможно последователността да е говор (speech possible sequence threshold), S_L е праг, при който е вероятно последователността да е говор (speech likely sequence threshold), L_S –таймер за кратко закъснение (short hangover time), and L_M – таймер за средно закъснение (medium hangover time).

В алгоритъма за фиксирани прагове (т. 3.4.1) се използват три параметъра, както следва: $\alpha = 0.03, \beta = 1.5$ и $\gamma = 0.05$.

В алгоритъма за адаптивните прагове (т. 3.4.2) се използват 6 параметъра, както следва: $\alpha_1 = 0.1; \beta_1 = 1.1; \alpha_2 = 0.05; \beta_2 = 1.2; \kappa = 0.5,$ и M=3.

Времевите константи, използвани в крайния автомат имат следните стойности в ms: *MaxQuietTime=2000; BegTime=300; MaxStateTime=1500; UpTime1=200; UpTime2=100; MiddleTime=200; MinLengthTime=500; EndTime=500*. Тези стойности са избрани така, че крайният автомат да може да обработва произнасяния с дължина по-голяма от *MinLengthTime*, но с паузи по-малки от *MaxStateTime*.

3.6.3. Определяне на точността при детекция

При тези експерименти точността на детекция се оценява чрез разликата (в брой сегменти) между автоматично (чрез предложените алгоритми) и ръчно определените гранични точки [Yamamoto et al., 2006]. За всяко произнасяне сегментната разлика (frames difference) $D_B(s)$ между ръчно и автоматично определените начални гранични точки е дефинирана като

$$D_B(s) = M_B(s) - ED_B(s), \tag{3.30}$$

където $M_B(s)$ е ръчно определената начална гранична точка; $ED_B(s)$ е началната гранична точка, определена чрез съответния алгоритъм и $s = 1,\ldots,S$ е броят на фразите. Сегментната разлика за крайните гранични точки $D_E(s)$ е определена като

$$D_E(s) = M_E(s) - ED_E(s), \tag{3.31}$$

където $M_E(s)$ е ръчно определената крайна гранична точка; $ED_E(s)$ е крайната гранична точка, определена чрез съответния алгоритъм.

Анализа на точността на детекция се осъществява чрез построяване на хистограми на сегментните разлики – поотделно за началните и крайните точки. Броят на използваните фрази е 262 и 84, съответно за BG-SRDat и TIDIGITS. Броят на интервалите (bins) в хистограмите е съответно 19 и 9. Тези стойности са получени като средни стойности на броя интервали, изчислени за всеки параметър чрез правилото на Scott [Scott, 2010]. Прието е да бъдат построени хистограми при диапазон на изменение на сегментните разлики - [-20; 20]. Всяка стойност в представените таблици показва в проценти вероятността за указаните диапазони (rate of distribution). Прието е, че значими при детекция диапазони на изменение на сегментните разлики са [-10;10] и [-5;5]. Разлики по-големи 10 сегмента (100 msec) в едната или другата посока в действителност са груби грешки, които в повечето случай влияят негативно върху процеса на разпознаване. Абсолютните стойности на сегментните разлики са означени в таблиците като $|D_B|$ и $|D_E|$, а с $\overline{D}$ е означена тяхната средна стойност.

По-горе в текста (т. 3.3) са разгледани линейни и логаритмични версии на различните признаци. Построяване на обща stacked хистограма например за всички логаритмични параметри обаче генерира графика, която трудно може да бъде визуално интерпретирана. Поради това е прието посочените по-горе параметри на хистограмите да бъдат представени за всички детектори в таблици. След което се избират детекторите, при които най-голяма част от разпределението на сегментните разлики е в приетите диапазони и за тях се визуализират съответните stacked хистограми. В текста са включени следните таблици:

- Таблица 3.4. – корпус BG-SRDat, адаптивни прагове (adapt2thr); логаритмична скала (log scale);
- Таблица 3.5. - корпус BG-SRDat, фиксирани прагове (fixed2thr); линейна скала (lin scale);
- Таблица 3.6. - корпус TIDIGITS, адаптивни прагове (adapt2thr); логаритмична скала (log scale);
- Таблица 3.7. - корпус TIDIGITS, фиксирани прагове (fixed2thr); линейна скала (lin scale);

Таблица 3.4. Вероятност в проценти

Speech Corpus	BG-SRDat					
Features & adapt2thr	$\|D_B\|$		$\|D_E\|$		$\bar{D}$	
	≤ 5	≤ 10	≤ 5	≤10	≤ 5	≤ 10
log-MTE-E	56.10	71.37	**55.72**	77.09	55.91	74.23
log-EE-E	49.61	65.26	36.64	58.01	43.12	61.64
log-MD-E	**60.30**	80.91	54.96	74.42	**57.63**	77.67
log-GDMD-E	54.96	**87.02**	51.90	**78.24**	53.43	**82.63**
LTSD-E	41.60	84.35	37.02	67.17	39.31	75.76
log-GDMD-H	47.32	**87.02**	35.11	61.06	41.22	74.04
LTSD-H	45.80	85.11	24.42	46.56	35.11	65.83

Таблица 3.5. Вероятност в проценти

Speech Corpus	BG-SRDat					
Features & fixed2thr	$\|D_B\|$		$\|D_E\|$		$\bar{D}$	
	≤ 5	≤ 10	≤ 5	≤10	≤ 5	≤ 10
MTE-E	37.78	53.05	**35.87**	55.72	36.83	54.38
EE-E	37.40	51.52	21.75	41.22	29.58	46.37
MD-E	41.22	66.41	25.57	45.41	33.39	55.91
lin-GDMD-E	**49.23**	**83.20**	35.11	**58.77**	**42.17**	**70.99**

Таблица 3.6. Вероятност в проценти

Speech Corpus	TIDIGITS					
Features & adapt2thr	$\|D_B\|$		$\|D_E\|$		$\bar{D}$	
	≤ 5	≤ 10	≤ 5	≤10	≤ 5	≤ 10
log-MTE-E	65.47	75.00	61.90	79.76	63.69	77.38
log-EE-E	60.71	77.38	66.66	84.52	63.69	80.95
log-MD-E	76.19	91.66	69.04	86.90	72.61	89.28
log-GDMD-E	85.71	97.61	67.85	**89.28**	**76.78**	**93.45**
LTSD-E	76.19	86.90	**76.19**	84.52	76.19	85.71
log-GDMD-H	**94.04**	**98.80**	58.33	80.95	76.19	89.88
LTSD-H	58.33	73.80	41.66	48.80	50.00	61.30

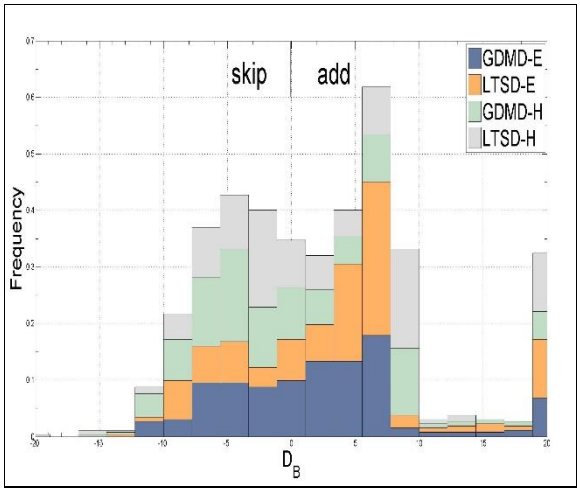
Таблица 3.7. Вероятност в проценти

Speech Corpus	TIDIGITS					
Features & fixed2thr	$\|D_B\|$		$\|D_E\|$		$\bar{D}$	
	≤ 5	≤ 10	≤ 5	≤10	≤ 5	≤ 10
MTE-E	27.38	54.76	40.47	55.95	33.92	55.35
EE-E	21.42	63.09	**46.42**	**67.85**	33.92	65.47
MD-E	33.33	80.95	33.33	53.57	33.33	67.26
lin-GDMD-E	**50.00**	**98.80**	44.04	61.90	**47.02**	**80.35**

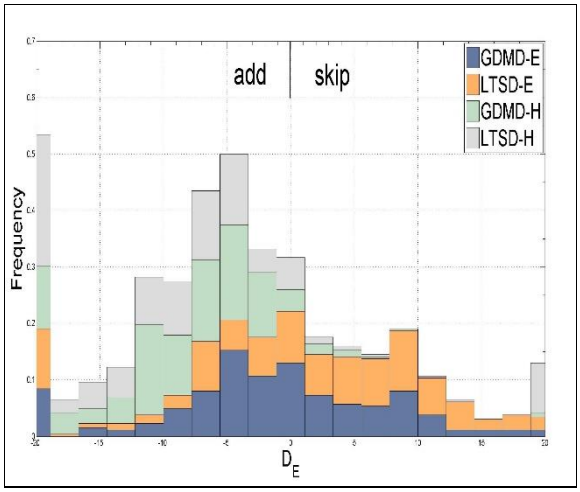
При анализ на резултатите в Таблици 3.4-3.7 е установено, че най-голямата вероятност сегментните разлики да попадат в анализираните диапазони се получава почти винаги при GDMD признака (виж средните стойности $\bar{D}$). Има само едно изключение и то е в таблица 3.4 – за MD признака. Най-висок среден процент (в колона $\bar{D}(\leq 10)$) при корпуса BG-SRDat е получен при log-GDMD-E & adapt2thr - 82.63%. При

корпуса TIDIGITS този процент (в колона $\bar{D}(\leq 10)$) е получен също за log-GDMD-E & adapt2thr – 93.45%. При по детайлен поглед например в таблица 3.4 за корпуса BG-SRDat и детектор log-GDMD-E & adapt2thr и сегментна разлика 10 - за началните гранични точки е получен резултат 87.02%, а са крайните -78.24%. Този резултат е много добър особено като се имат предвид фонемните особености на използваната фраза и зашумения телефонен сигнал.

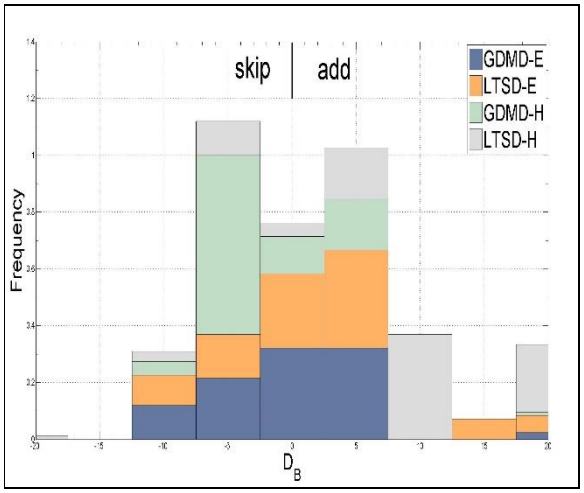
Като цяло (от гледна точка на средните стойности в приетите диапазони) най-добри резултати са получени при признаци в логаритмичната скала, съчетани с алгоритъма за адаптивни прагове. Избират се следните четири детектора: log-GDMD-E & adapt2thr, log-GDMD-H & adapt2thr, LTSD-E & adapt2thr и LTSD-H описан в [Luengo et al., 2010, §2] и за тях е построена обща stacked хистограма. На фиг. 3.18 и 3.19 са показани хистограмите на сегментните разлики D_B и D_E получени за данни от корпуса BG-SRDat, а на фиг. 3.20 и 3.21 - за данни от корпуса TIDIGITS.



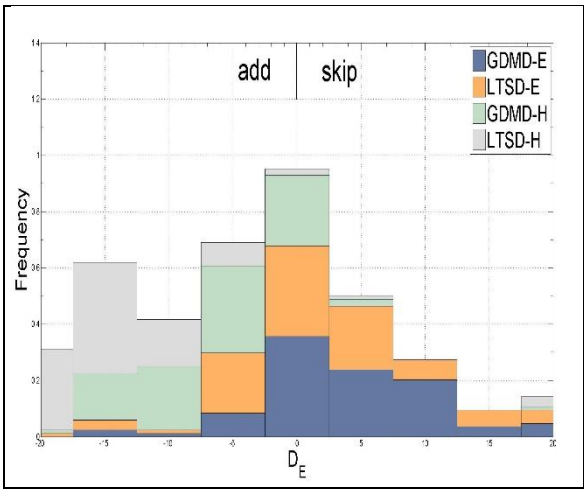
Фиг. 3.18. Хистограми на D_B-корпус BG-SRDat.



Фиг. 3.19. Хистограми на D_E-корпус BG-SRDat.



Фиг. 3.20. Хистограми на D_B-корпус TIDIGITS.



Фиг. 3.21. Хистограми на D_E-корпус TIDIGITS.

На хистограмите представени във фигури 3.18 - 3.21 са добавени два етикета "*skip*" (пропуснати) и "*add*" (добавени). Тези етикети обозначават области в хистограмите които съответстват на пропуснати и добавени сегменти.

За фразата от корпуса BG-SRDat, както се вижда на фиг. 3.18, хистограмата на сегментните разлики D_B при началните гранични точки е двумодална. Това се дължи на факта, че за някои записи всички детектори пропускат фонемата *'з'* и поставят началната точка на фразата в началото на звучната съгласна *'д'*. Тези грешки съответстват на лявата мода на разпределението и тя има стойност на разликите около [-5] сегмента. Дясната мода на разпределението (около [+5] сегмента) е резултат от добавени шумови сегменти преди първата фонема *'з'* което е резултат от свойствата на логаритмичната функция да усилва стойностите на контура с ниско ниво. В текста не е показана хистограмата на параметрите в линейната скала от таблица 3.7, но в [Ouzounov, 2014] е включена и там се вижда, че е налице значително количество пропуснати сегменти - лявата мода на разпределението има стойност около [-10], докато дясната мода е около стойност [0], т.е. добавени са минимално количество сегменти с шум преди първата фонема. Фразата от BG-SRDat завършва с шумоподобната съгласна *'с'* която е трудно да се локализира в зашумена среда. В този случай в хистограмата на сегментните разлики D_E за крайните гранични точки показана на фиг. 3.19 има един максимум и той е при стойност на разликите около [-5]. Или е налице предимно добавяне на шумови сегменти след края на фразата. Значителна стойност в разпределението съществува при сегментна разлика около [-20], т.е. добавени са шумови фрагменти с дължина приблизително 200 ms. И при четирите детектора се наблюдава подобна грешка като при LTSD-H тя е най-голяма.

Говорните данни от корпуса TIDIGITS са последователности от цифри т.е. граничните фонеми са различни и това води до значително по-добри резултати при всички детектори (с изключение на LTSD-H) – сравняват се стойностите в колона $\bar{D}(\leq 10)$ при таблици 3.4 и 3.6 и таблици 3.5 и 3.7. На хистограмите във фиг. 3.20 и фиг. 3.21 не се наблюдават значителни стойности (в сравнение с корпуса BG-SRDat) на разпределението при сегментни разлики [±20]. Доколкото ги има тези стойности са получени предимно от детектора LTSD-H.

Важно е да се сравнят резултатите получени от два различни детектора които използват еднакви параметри и прагов механизъм, но в единия от тях се прилага краен автомат за определяне на граничните точки, докато в другия – hangover алгоритъм. Това са двойките: (log-GDMD-E & adapt2thr vs log-GDMD-H & adapt2thr) и (LTSD-E &

adapt2thr vs LTSD-H). За корпуса BG-SRDat в таблица 3.4 стойностите в колона $\bar{D}(\leq 10)$ за съответните двойки са: (82.63 vs 74.04) и (75.76 vs 65.83). За корпуса TIDIGITS в таблица 3.6 стойностите в колона $\bar{D}(\leq 10)$ за съответните двойки са: (93.45 vs 89.88) и (85.71 vs 61.30). На базата на тези числени стойности може да се твърди, че чрез предложения краен автомат се постига по-висока точност на детекция на граничните точки в сравнение с hangover алгоритъма.

Или казано с други думи, ако фразата започва или завършва с шумови съгласни, детектора, който използва hangover алгоритъма [ETSI, 2007] удължава фразата повече от детектора, който използва предложения краен автомат. Този факт е резултат от различния начин на работа на двата алгоритъма. При крайния автомат се формира множество от пресечни точки на контура с два прага. Граничните точки се избират от това множество чрез набор от логически правила. Тези правила използват времеви периоди, които въвеждат ограничения в изменението на контура и тези периоди са различни за началото и за края на фразата. При hangover алгоритъма се забавя прехода в края на фразата с цел да се прихванат шумовите съгласни, ако има такива. Анализира се буфер с фиксирана дължина съдържащ вече приетите решения 0/1 (шум/говор) които са получени чрез сравнение с единичен праг (в повечето случаи). Максималния брой последователни единици разположени след текущия сегмент се сравнява с предефинирани времеви периоди. След това сравнение окончателната бинарна последователност може да бъде променена в зависимост от вида на съседните сегменти.

3.6.4. Зависима от текста верификация на диктори

В тази част от работата всеки един от избраните в т. 3.5.3 детектори на гранични точки е включен в система за верификация на диктори. Целта на подобно изследване е да се оцени ефективността на съответния детектор чрез неговото влияние върху грешката при разпознаване. Допълнително е осъществена тази верификация и с ръчно сегментирани говорни данни.

Използвани са два различни подхода при верификация на диктори - DTW и HMM. Тестовете са реализиран с кратки фрази на български език записани по телефонен канал и избрани от корпуса BG-SRDat. По-долу в текста е представено кратко описание на двете системи за верификация [SR-API, 2003].

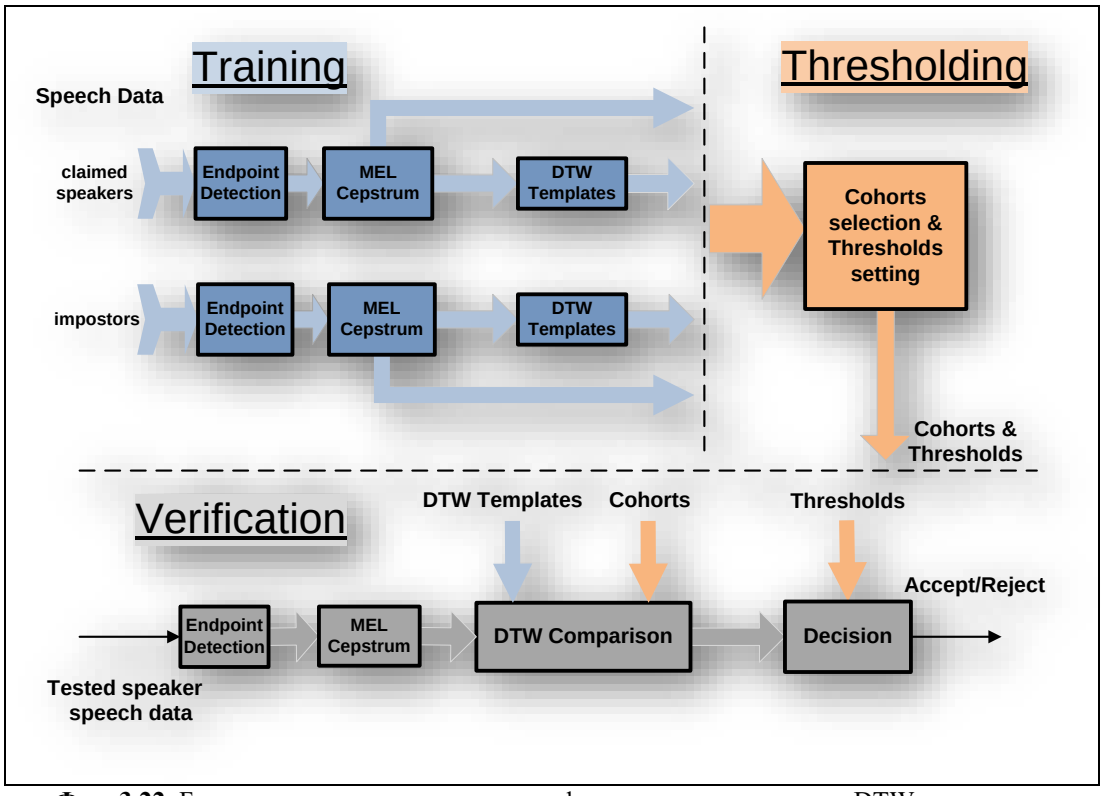
3.6.4.1. Предварителна обработка

Върху говорния сигнал се прилага тегловна функция на Хеминг с дължина 30 ms и изместване 10 ms. Като параметрични представяне се използва Мел-кепстъра (Mel-

Frequency Cepstral Coefficients - MFCC) с размерност 14 (без нулевия коефициент) получен чрез 24 Мел-филтри с еднаква площ (equal-area filter-bank). За всеки файл се прилага CMS [Ganchev, 2011].

3.6.4.2. Верификация на диктори чрез DTW

Използван е normalize-wrap DTW алгоритъм с локални ограничения на Itakura [Theodoridis et al., 2010] и кепстрално разстояние. Еталонната и тестовата параметрични последователности се нормализират по дължина преди да се приложи DTW алгоритъма. Сравнението на последователностите се извършва при ограничени гранични условия (constrained boundary conditions). Еталона на даден диктор (template) е получен чрез усредняване (след DTW подравняване - alignment) на предназначените за обучение параметрични последователности. Индивидуалните прагове за верификация са определени чрез кохортна нормализация (cohort normalization) [Tistarelli et al., 2014]. Блоковата схема на използваната в работата система за верификация на диктори чрез DTW алгоритъма е показана на фиг. 3.22.

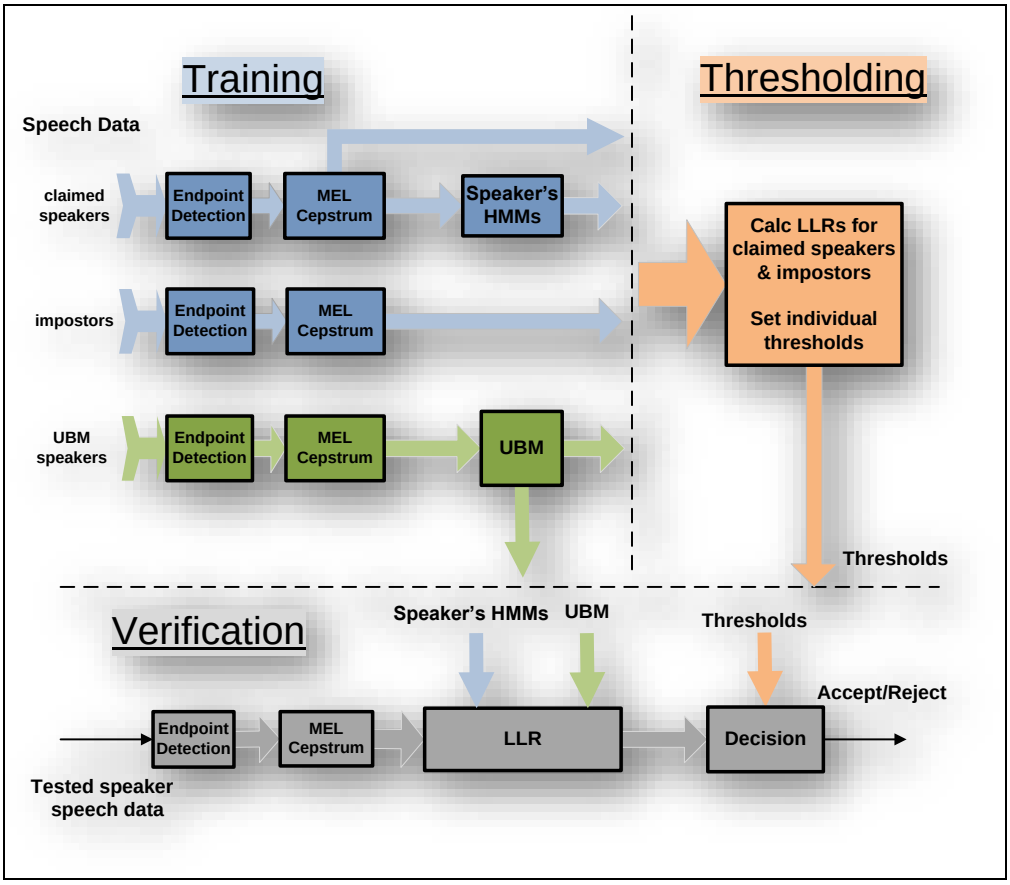


Фиг. 3.22. Блокова схема на система за верификация на диктори чрез DTW алгоритъм.

3.6.4.3. Верификация на диктори чрез HMM

Използван е whole-phrase HMM, при който анализираната фраза е моделирана като цяла последователност [Buyuk at al., 2012]. Избран е модел с топология ‘left-to-right‘, no skip state и разпределенията се моделират като смес от Гаусови разпределения с диагонални ковариационни матрици. Обучението на модела е реализирано с алгоритъма на Baum-Welch [Gales, et al., 2008].

При верификацията на дикторите са използвани индивидуални прагове. Те са изчислени чрез т.н. фонов модел (world/background model) описващ глобалното множество от диктори различни от анализираните диктори. Числената оценка за всеки диктор е получена чрез изчисляване на логаритмичното отношение на правдоподобие за дадено произнасяне чрез използване на модела на диктора и фоновия модел. Праговете се установяват a priori на базата на разпределенията на числените оценки на целевите (claimed/target speakers) и нецелевите диктори (impostors) [Munteanu et al., 2010]. Блоковата схема на използваната в работата система за верификация на диктори чрез HMM алгоритъма е показана на фиг. 3.23.



Фиг. 3.23. Блокова схема на система за верификация на диктори чрез HMM алгоритъм.

3.6.4.4. Данни, използвани при верификация

Данните, използвани при експериментите са избрани от корпуса BG-SRDat и включват 337 записа (произнасяния), събрани от 18 диктора (мъже). По-голямата част от тях – 262 записа от 12 диктора (тези данни са едни и същи и при двете разпознаващи системи) са предназначени за създаване на еталони (DTW) или модел (HMM) на диктора (training set), за установяване на прагове (validation set) и за тестване (verification set). Тъй като корпуса е с недостатъчно количество данни (small set) едни и същи данни се използват за обучение и за установяване на праговете [Bengio et al., 2004]. Останалите данни – 75 записа от 6 диктора са избрани, за да формират фоновия модел (universal background model-UBM) при HMM верификация.

За да се използват ефективно всичките разполагаеми данни е използван подхода с 5x2 cross validation [Kuncheva, 2014]. Крайните резултати в случая са изчислени като претеглени средни стойности на резултатите от петте повторения. При всеки единичен тест в режим на верификация има 142 теста за фалшиво отхвърляне (False Rejection tests - FR) и 1562 теста за фалшиво приемане (False Acceptance tests - FA). След 5 повторения при 5x2 cross validation общия брой на тестовете е: за фалшиво отхвърляне – 710 и за фалшиво приемане – 7810.

3.6.4.5. Експериментални резултати

Ефективността на отделните алгоритми за определяне на граничните точки е сравнена чрез грешката, получена при верификация на диктори. Допълнително е извършена верификация със същите произнасяния, но при тях граничните точки са определени ръчно.

Известно е, че при корпуси с малък брой данни и освен това записани в реална среда (real - world data) грешката от верификация не е достатъчно надеждна оценка за качествата на дадена разпознаваща система [Bengio et al., 2004]. Тъй като това е точно разглежданата в работата ситуация бе решено да се приложи методологията за оценка на резултатите получени при верификация на диктори предложена в [Bengio et al., 2004].

Резултатите от верификацията на дикторите се представят като отношения - False Rejection Rate (FRR), False Acceptance Rate (FAR) и Half Total Error Rate (HTER). Те са дефинирани във вида:

$$FRR = \frac{FR}{NC};$$
$$FAR = \frac{FA}{NI};$$
$$\text{HTER} = \frac{FRR + FAR}{2}, \tag{3.32}$$

където *FA* е общият брой фалшиви приемания направени от системата, *FR* е общият брой фалшиви отхвърляния, *NC* е броят на тестовете с целеви диктори (client/target speaker accesses), а *NI* е броят на тестовете с нецелеви диктори (impostor/non-target speaker accesses).

Също така съгласно [Bengio et al., 2004] е изчислен и 95% доверителен интервал - Confidence Interval (CI) –за HTER. Z_{HTER} -теста, предложен в [Bengio et al., 2004] е използван, за да се провери доколко резултатите от даден класификатор са статистически значимо различни от тези на друг. Когато се сравняват резултати от системи за разпознаване, които включват много идентични модули и данните за обучение и тестване са еднакви то трябва да се предполага, че получените резултати са силно корелирани [Bengio et al., 2004]. Имайки предвид този факт както и коментарите в [Bengio et al., 2004], в работата е решено да се използва версията на Z_{HTER} -теста, при която няма оптимистично изместване на резултата.

В текста по-долу за яснота някои от изреченията са опростени, например ‘при детектора GDMD-E се получава по-малка грешка при верификация‘ означава, че при системата за верификация, която използва този детектора се получава по-малка грешка, а не че самия детектор генерира по-малка грешка.

Резултатите, получени при верификация на диктори чрез DTW алгоритъма са показани, както следва: в таблица 3.8 - грешките FRR, FAR и HTER в проценти и стойностите на 95% доверителен интервал; в таблица 3.9 – доверителната вероятност в проценти за всяка двойка сравнявани детектори.

Таблица 3.8. DTW – Грешки при верификация на диктори

Endpoint Detector	FRR, %	FAR, %	HTER, %	95%CI
Manual	7.88	4.83	6.35	±0.0101
GDMD-E	9.43	8.22	8.82	±0.0111
LTSD-E	10.70	11.57	11.13	±0.0119
GDMD-H	10.84	6.42	8.63	±0.0117
LTSD-H	11.83	11.65	11.74	±0.0124

Таблица 3.9. DTW тест
Доверителна вероятност в проценти за всяка двойка детектори

Endpoint Detectors	Manual	GDMD-E	LTSD-E	GDMD-H	LTSD-H
Manual	–	99.86	100.00	99.58	100.00
GDMD-E	–	–	99.41	18.63	99.93
LTSD-E	–	–	–	99.66	50.96
GDMD-H	–	–	–	–	99.96
LTSD-H	–	–	–	–	–

Известно е, че броят на състоянията при HMM се избира емпирично, но е препоръчително те да бъдат пропорционални на броят на фонемите в анализираната фраза [Buyuk at al., 2012]. Използваната фраза на български език съдържа 6 различни думи включващи общо 31 фонеми – 10 гласни и 21 съгласни. Фразата съдържа също и 5 паузи между думите. С цел да се определи подходяща топология за използвания Марковски модел е реализирана верификация на диктори с ръчно сегментирани фрази и различен брой на състояния и компоненти (mixtures) в Гаусовите смеси. Броят на състоянията е 18, 25, 35, а броят на компоненти е 2, 3, 4. Минимална грешка при верификация (HTER=8.42%) е получена за модел с 35 състояния и брой компоненти – 2 и тази топология е използвана при всички останали експерименти.

Резултатите, получени при верификация на диктори чрез HMM алгоритъма са показани, както следва: в таблица 3.10 - грешките FRR, FAR и HTER в проценти и стойностите на 95% доверителен интервал; в таблица 3.11 – доверителната вероятност в проценти за всяка двойка сравнявани детектори.

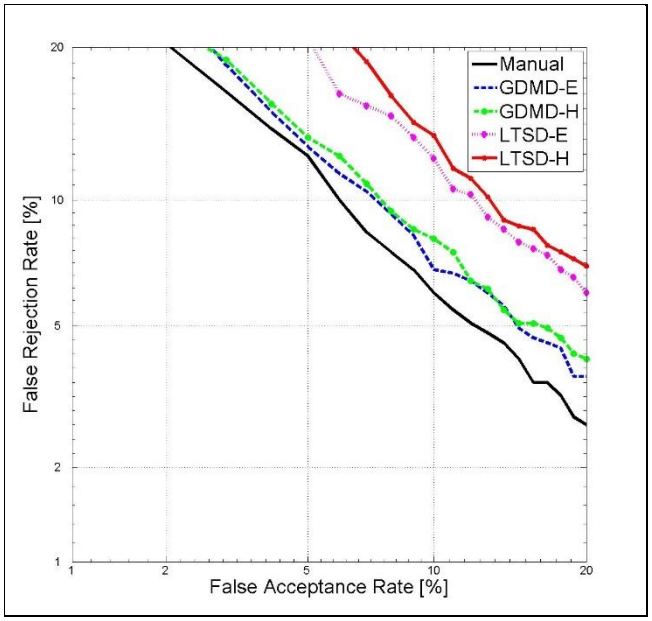
Таблица 3.10. HMM – Грешки при верификация на диктори

Endpoint detector	FRR, %	FAR, %	HTER, %	95%CI
Manual	15.63	1.21	8.42	±0.0134
GDMD-E	18.45	0.98	9.71	±0.0143
LTSD-E	22.25	1.20	11.72	±0.0153
GDMD-H	18.45	1.02	9.73	±0.0143
LTSD-H	22.53	1.04	11.78	±0.0154

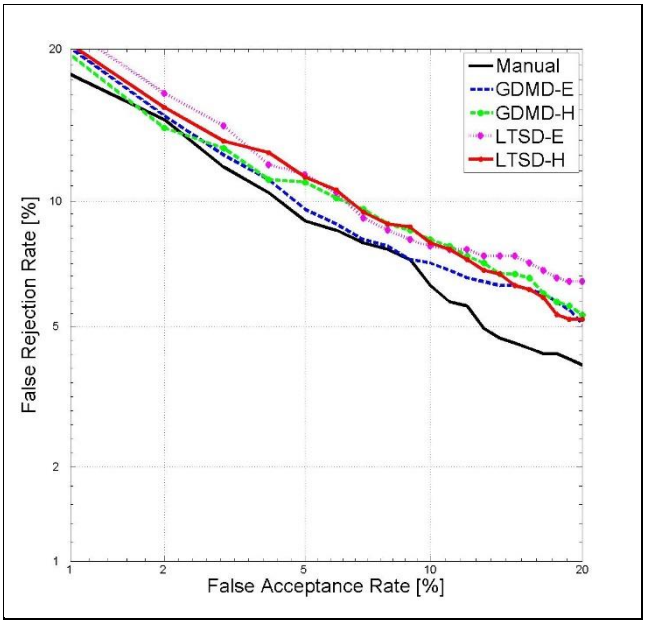
Таблица 3.11. HMM тест
Доверителна вероятност в проценти за всяка двойка детектори

Endpoint Detectors	Manual	GDMD-E	LTSD-E	GDMD-H	LTSD-H
Manual	–	80.44	99.85	81.12	99.87
GDMD-E	–	–	93.95	1.54	94.63
LTSD-E	–	–	–	93.69	4.31
GDMD-H	–	–	–	–	94.39
LTSD-H	–	–	–	–	–

На фиг. 3.24 и фиг. 3.25 са показани усреднените DET графики [Beigi, 2011] визуализиращи изменението на грешката при верификация получена за всеки един детекторите на гранични точки.



Фиг. 3.24. DTW - DET графики за различни детектори.



Фиг. 3.25. HMM - DET графики за различни детектори.

3.6.4.6. Коментари

На фиг. 3.24 и фиг. 3.25 ясно се вижда, че графиките на детекторите използващи GDMD признака са по-близо до еталонната графика (верификация чрез ръчно определени гранични точки) в сравнение с всички останали детектори. Тези резултати са в съответствие с резултатите представени в таблици 3.8 и 3.10, т.е., при детекторите

GDMD-E и GDMD-H се получава минимална грешка при верификация. От двата детектора GDMD-E е по-добър от гледна точка на точността при детекция и за двата корпуса с данни (таблици 3.4 и 3.6) и при HMM теста (таблица 3.10), докато вторият детектор е незначително по-добър в DTW теста (таблица 3.8). Въз основа на Z_{HTER} - теста може да се твърди, че двата детектора са равностойни (equivalent). По-големите грешки при верификация на диктори получени за детекторите LTSD-E и LTSD-H са резултат от наличието на съществени грешки при определяне на граничните точки. За тези детектори на фигури 3.18 - 3.21 се наблюдават голямо количество сегментни разлики, които са по-големи от 20 сегмента (200 ms).

При подробен анализ на експерименталните резултати са установени следните факти:

- във всички тестове при детекторите използващи log-GDMD признака се получават по-добри резултати от тези при детектори на базата на LTSD;
- при детектора GDMD-E, точността на детекция е максимална и за двата корпуса с данни;
- при верификация на диктори най-малка грешка е получена при GDMD-H детектора за DTW теста и при GDMD-E детектора за HMM теста;
- Z_{HTER}-теста показва, че при GDMD детекторите резултатите от разпознаване са статистически значимо различни от тези при LTSD детекторите. При DTW и HMM тестове за двойките детектори [GDMD-E, LTSD-E]; [GDMD-E, LTSD-H]; [GDMD-H, LTSD-E]; [GDMD-H, LTSD-H] се получава доверителна вероятност по-голяма от 90%;
- Z_{HTER}-теста показва, че при всички анализирани детектори по-значима част (в статистически смисъл) от тях е признакът, а не алгоритъма за взимане на решение (decision scheme). Двойките детектори с еднакви признаци, но с различни алгоритми за взимане на решение са равностойни (equivalent). Това е валидно за двойките детектори [GDMD-E, GDMD-H] и [LTSD-H, LTSD-E] при които се получава много ниска стойност на доверителната вероятност.

3.7. Заключение

В настоящата глава е извършен сравнителен експериментален анализ на ефективността на предложените в гл. 2 признаци предназначени за детекция на говор чрез анализ на времеви контури. Оценката на ефективността е реализирана в три етапа. На първия етап за даден признак е определено Евклидово разстояние между два Z-нормализирани

времеви контура получени съответно от чист и зашумен запис. По този начин е налице количествена оценка на структурните различия на контурите предизвикани от влиянието на даден вид шум върху свойствата на конкретен признак. На втория етап е предложен подход за разработване на детектор на гранични точки включващ алгоритъм за изчисляване на адаптивни прагове и детерминиран краен автомат. На базата на този подход и в зависимост от характеристиките на контурите на признаците са разработени три алгоритъма за определяне на гранични точки. Тук тяхната точност е оценена чрез анализ на разликите между ръчно определените и получените от съответния алгоритъм гранични точки. На третия етап е оценено влиянието на алгоритмите за ОГТ върху точността на разпознаване в две системи за зависима от текста верификация на диктори базирани съответно на DTW и HMM алгоритмите.

На базата на получените експериментални резултати са направени следните три заключения:

- Първо – детекторите на базата на log-GDMD признака във всички тестове превъзхождат тези на базата на LTSD. Необходимо е да се уточни, че LTSD признакът е адаптивен спрямо променящото се ниво на шума във фразата, докато при log-GDMD се разчита само на присъщата (вътрешна) робастност на неговите два компонента – на модифицирания спектър на групово закъснение и на делта спектралната автокорелационна функция;
- Второ –точността на детекция при използване на краен автомат с адаптивни прагове винаги превъзхожда, при един и същи признак, тази получена чрез hangover алгоритъма;
- Трето – от гледна точка на грешката при верификация в повечето случаи детектора с краен автомат и адаптивни прагове превъзхожда, при един и същ признак, този с hangover алгоритъма, но разликата между тях не статистически значима.

3.8. Резюме на получените резултати към Глава 3

Научни резултати:

1. Предложен е подход за определяне на гранични точки на говорно съобщение включващ алгоритъм за изчисляване на адаптивни прагови стойности и детерминиран краен автомат (Глава 3, т. 3.4.1-3).

Научно-приложни резултати:

1. Направен е сравнителен експериментален анализ на предложените в гл. 2 признаци спрямо избрани референтни такива. Сравнението е на базата на Евклидовото разстояние между Z-нормализирани времеви контури, изчислени за всеки признак съответно от чист и от зашумен сигнал (Глава 3, т. 3.3).
2. Разработени са три алгоритъма за определяне на гранични точки базиращи се на предложения подход и формирани съобразно използваните времеви контури. Оценена е тяхната точност чрез хистограмен анализ на разликите между ръчно определените и получените от съответния алгоритъм гранични точки. Експериментите са реализирани със зашумени говорни данни на български и английски език (Глава 3, т. 3.5 и т. 3.6.3).
3. Направен е сравнителен експериментален анализ на ефективността на предложените в гл. 2 признаци спрямо избрани референтни такива при използването им в предложените алгоритми за определяне на гранични точки. Сравнението е на базата на грешката при разпознаване в две системи за зависима от текста верификация на диктори базирани съответно на DTW и HMM алгоритми. При експериментите са използвани говорни данни на български език записани по телефонен канал (Глава 3, т.3.6.4).

ГЛАВА 4

Алгоритми за детекция на говор при независима от текста идентификация на диктори. Експериментално изследване.

4.1. Увод

Откриването на говорни сегменти в даден аудио сигнал се нарича Voice Activity Detection - VAD. По принцип този тип детекция е реализирана чрез бинарна класификация. Независимо от широкото използване на VAD-алгоритми засега не съществува универсален алгоритъм, който да работи надеждно при различни задачи и в реална среда (real-world environment). Това се дължи на изискванията за всеобщност (необходимост от прилагане във всички системи за обработка на говор) и различната степен на сложност (зависи от конкретните цели на дадена система).

По-детайлно описание на VAD-алгоритмите е представено в обзорната част на дисертацията поместена в т. 1.2.

В настоящата глава е извършен сравнителен експериментален анализ на ефективността на предложените в гл. 2 признаци предназначени за детекция на говор. За всеки признак (референтен или предложен от автора) се формира отделен детектор на говор, който става част от система за независима от текста идентификация на диктори реализирана чрез многослоен перцептрон (МСП).

Експерименталните изследвания са реализирани с два различни алгоритма за детекция на говор – те условно ще бъдат означени в текста като VAD-1 и VAD-2. Разработването на два отделни VAD модула се налага поради факта, че в гл. 2 са предложени два вида признаци. Едните са скаларни величини и са предназначени за детекция на говор чрез контурен анализ, докато другите са вектори и се използват като параметрични представяния в класификационни схеми.

Във VAD-1 като класификатор е използван МСП и бинарното решение се получава чрез прагов алгоритъм анализиращ стойностите, получени в изходните неврони на перцептрона. При VAD-2 се използва времеви контур и прагове (подобно на алгоритмите разгледани в гл. 3). За всеки сегмент стойността на контура се сравнява с фиксиран или адаптивен праг и се приема решение не-говор/говор. Впоследствие,

аналогично на VAD-1, само говорните сегменти се подават към алгоритъма за разпознаване [Kitaoka et al., 2007].

Като референтни параметри за VAD-1 са избрани многолентова спектрална ентропия (Multi-Band Spectral Entropy - MBSE) [Misra et al., 2005]; признак, получен чрез филтрация в честотната област (Frequency-Filtering parameter - FF) [Macho et al., 2001]; относителна спектрална разлика (Relative spectral difference - RSD) [Macho et al., 2001] и Мел-кепстър с индексен лифтер (index-weighted Mel-Frequency Cepstral Coefficients – IW-MFCC) [Ganchev, 2011].

При VAD-2 референтни параметри са съответно получените чрез алгоритъма на Sohn [Sohn et al., 1999] (разгледан в гл. 1. - т. 1.2.3.1.1) - тук е използвана VoiceBox-версията на алгоритъма [VoiceBox, online]; алгоритъма на Wu [Wu et al., 2006] и разгледания в гл. 3 (т. 3.2.4) алгоритъм LTSD [Ramirez et al., 2004].

Оценката на ефективността на предложените параметри е реализирана на два етапа. На първия етап е оценена точността на различните детектори чрез анализ на разликите между ръчно определените и получените от съответния алгоритъм гранични точки на говорните фрагменти. Тестове са реализирани с говорни данни на английски език, избрани от корпусите – TIDIGITS [Dan Ellis, online] и NOIZEUS [NOIZEUS, online] и на български език - от корпуса BG-SRDat [виж гл. 5] и [Ouzounov, 2003].

На втория етап е оценено влиянието на различните VAD-алгоритми върху точността на разпознаване в система за независима от текста идентификация на диктори, при която се използва МСП с един скрит слой (Multi-Layer Perceptron – MLP). Тук тестовете са реализирани с фрази на български език записани по телефонен канал и избрани от корпуса BG-SRDat.

С цел постигане на по-голяма яснота в изложението двата детектора VAD-1 и VAD-2 ще бъдат разгледани поотделно – структура, параметри, използвани данни и получени експериментални резултати. След което получените резултати са анализирани и обобщени.

4.2. Референтни признаци

4.2.1. Алгоритъм на Wu

Говорният сигнал в описания алгоритъм [Wu et al., 2006] се разлага в четири честотни ленти чрез прилагане на дискретно уейвлет преобразуване (Discrete wavelet transform - DWT). Във всяка лента чрез коефициентите на DWT се изчислява параметъра Speech Activity Envelope (SAE). Той се получава чрез последователно изпълнение на следните

преобразувания - енергийния оператор на Teager, автокорелационна функция (sub-band signal auto-correlation function - SSACF) и среден-делта признак [Ouzounov, 2004] означен като Delta Sub-band Signal Auto-Correlation Function (DSSACF). Експерименталните резултати показват, че чрез този признак се постига успешна детекция на говор в зашумена среда като особено ефективна е тя при шум с променливо ниво.

За *n-я* сегмент и за дадено ниво j DWT разлага сигнала в $j+1$ честотни ленти, съответстващи на множеството от уейвлет коефициенти $w_{k,n}^{j}$, в конкретния случай *j=3* или

$$w_{k,m}^{3}(n) = DWT\left\{s(n,l),3\right\},\ l=1,...,L;\ k-1,...,4, \tag{4.1}$$

където с $w_{k,m}^{3}$ е означен *m-я* коефициент в *k-mama* честотна лента, а L е дължината на сегмента. Дължината на декомпозирания сигнал във всяка лента е $L/2^{k}$.

След прилагане на енергийния оператор на Teager $\psi_d[.]$ върху уейвлет коефициентите се получава

$$t_{k,m}^{3}(n) = \psi_d[w_{k,m}^{3}(n)], k=1,...,4. \tag{4.2}$$

Автокорелационната функция SSACF е

$$R_{k,m}^{3}(n) = R[t_{k,m}^{3}(n)], \tag{4.3}$$

където *R[.]* е автокорелационен оператор. Делта функцията DSSACF има вида

$$\widehat{R}_{k,m}^{3}(n) = \Delta[R_{k,m}^{3}(n)], \tag{4.4}$$

където с Δ е означено разгледаното в [Ouzounov, 2004] делта преобразуване. Параметъра MDSSACF се получава като средна стойност или

$$\overline{R}_{k}^{3}(n) = E[\widehat{R}_{k,m}^{3}(n)], \tag{4.5}$$

Крайния параметър SAE има вида

$$SAE(n) = \sum_{k=1}^{4} \overline{R}_{k}^{3}(n). \tag{4.6}$$

4.2.2. Многолентова спектрална ентропия

При определяне на спектралната ентропия за *n-я* сегмент първо се изчислява функцията на вероятността (Probability Mass Function – PMF) $P(\left|X(n,k)\right|^2)$ за спектъра на мощността (за целия честотен диапазон) $\left|X(n,k)\right|^2$ и съгласно [Misra et al., 2005] тя е

$$P(\left|X(n,k)\right|^2) = \frac{\left|X(n,k)\right|^2}{\sum_{l=0}^{K/2}\left|X(n,l)\right|^2}, \tag{4.7}$$

където $k = 0,...,K/2$, K е размера на FFT и $n = 0,...,N-1$, N е броят сегменти.

Спектралната ентропия $H(n)$ за *n-я* сегмент е изчислена както следва

$$H(n) = -\sum_{k=0}^{K/2} P(|X(n,k)|^2).log_2(P(|X(n,k)|^2)) \quad (4.8)$$

Ентропията в (4.8) се разглежда като ентропия за целия честотен спектър (full-band spectral entropy). За да е възможно чрез спектралната ентропия да се получи по-детайлно описание за измененията в кратковременния спектър (например за местоположението и нивото на формантите) в [Misra et al., 2005] е предложено да се въведе многолентова спектрална ентропия. При нея спектъра се разделя на честотни ленти и ентропията се изчислява за всяка лента като се използва full-band PMF.

Съгласно [Misra et al., 2005] вектора на многолентовата спектрална ентропия - Multi-Band Spectral Entropy - MBSE за *n-я* сегмент има вида $\{H_{MBSE}(n,1),\ldots,H_{MBSE}(n,G)\}$ като неговите компоненти се изчисляват съгласно

$$H_{MBSE}(n,g) = -\sum_{k=B_g}^{B_{g+1}} P(|X(n,k)|^2).log_2(P(|X(n,k)|^2)) \quad (4.9)$$

където $P(|X(n,k)|^2)$ е full-band PMF в (4.7); $g = 1,\ldots,G$, G е броят честотни ленти и $\{B_1,B_2\}\cdots\{B_g,B_{g+1}\}\cdots\{B_{2G-1},B_{2G}\}$ са двойките гранични спектрални точки за всяка честотна лента.

4.2.3. Параметри на основата на спектрални производни

Ако $E(\omega)$ е обвивката на спектъра на анализирания говорен сигнал и $S(\omega) = \log(E(\omega))$, то производната на $S(\omega)$ спрямо честотата ω е

$$\frac{d}{d\omega}S(\omega) = \frac{1}{E(\omega)}\frac{d}{d\omega}E(\omega) \quad (4.10)$$

Когато енергията на спектъра е дефинирана в дискретна честотна скала (например получена чрез набор от лентови филтри) първата производна в (4.10) може да се представи чрез разлика. Ако наклона на спектъра е дефиниран като разлика между двата отчета заобикалящи текущия, то производната на логаритмичния спектър се представя като разлика между стойности на логаритмичния спектър или това е FF (Frequency-filtered band energies) параметъра [Macho et al., 2001]. Той се дефинира във вида

$$S_{FF}(k) = S(k+1) - S(k-1) \quad (4.11),$$

където k е индексът на съответната честотна лента. Този параметър може да се разглежда като получен чрез филтрация в честотната област на спектралните енергии с филтър от вида $z - z^{-1}$ [Macho et al., 2001].

Аналогично относителната спектрална производна може да се представи като относителна спектрална разлика (Relative Spectral Difference) във вида

$$S_{RSD}(k) = \frac{E(k+1) - E(k-1)}{E(k)} \tag{4.12}$$

За да се избегнат случаите с наличие на малки стойности в знаменателя се въвежда локална средна стойност на спектралните компоненти, получени от лентовите филтри и ф-ла (4.12) приема вида

$$S_{RSD}(k) = \frac{E(k+1) - E(k-1)}{E(k+1) + E(k) + E(k-1)} \tag{4.13}$$

В [Padrell et al., 2005] FF параметъра и линеен дискриминантен анализ са използвани успешно при детекция на говор.

4.2.4. МЕЛ-кепстър

МЕЛ-кепстърът е получен от спектър, чиято честотна скала е преобразувана в МЕЛ-скала. Най-разпространената версия на МЕЛ-кепстъра е тази, при която FFT спектъра се филтрира чрез набор от лентови филтри, разположени по честота съгласно критичните области на човешкия слух [Ganchev, 2011]. Определят се енергиите на изхода на филтрите и чрез прилагане на обратно косинус преобразуване върху логаритмичните стойности на енергиите се получава МЕЛ-кепстъра. Ако логаритмичната енергия на изхода на k-я филтър е означена с $\log(E(\mathrm{k}))$, то МЕЛ-кепстъра $c_{mel}(m)$ е

$$c_{mel}(m) = \sum_{k=1}^{K} \log(E(\mathrm{k})) \cos(m(k-0.5)\pi/K), \tag{4.14}$$

където K е броят на лентовите филтри; $m = 0,...,M$ и M е броят на кепстралните коефициенти.

Известно е, че вариативността на кепстралните коефициенти с малък пореден номер е резултат предимно от промените на предавателната характеристика на комуникационната среда. Докато измененията в коефициентите с висок пореден номер се предизвикват предимно от влияние на гласовия източник и др. [Ganchev, 2011].

За да се ограничи в кепстъра влиянието на едни или други спектрални изменения, най-често се прилага лифтер с определена форма. Тук е прието да бъде използван индексен лифтер – при него кепстралните коефициенти се умножават с поредния си номер. Основната причина за този избор са резултатите от разпознаване на диктори

представени в [Ouzounov, 2010]. В работата е посочено, че най-ефективна (т.е. при нея се получава минимална грешка) е комбинацията от root-power-sum кепстрално разстояние (т.е. кепстър с индексен лифтер) и МЕЛ-кепстър получен чрез лентови филтри с еднаква площ (equal-area filter-bank). Имено тази комбинация е използвана при реализираните тук изследвания.

4.3. Грешки при детекция

Точността на детекция се определя чрез сравняване на ръчно определените гранични точки с тези получени от съответния детектор. При нейната оценка се използват множество от грешки всяка от които описва различни характеристики на VAD алгоритъма. Например грешките условно означени като *изрязване* (clipping) показват каква част от говорния фрагмент грешно се класифицира като шум. И обратно, грешки означени като *добавяне* (insertion) показват каква част от шумовия сегмент грешно се класифицира като говор. Често използваните грешки са описани в [Davis et al., 2006]. Те са:

- *Front-End Clipping (FEC)*

При преход от шум към говор част от говорните сегменти грешно се класифицират като шум. Дефинира се във вида:

$$FEC = \frac{N_F}{N_{SPEECH}}, \tag{4.15}$$

където N_F е брой на говорните сегменти грешно класифицирани като шум при преход от шум към говор и N_{SPEECH} е общ брой говорни сегменти, получени при ръчна детекция.

- *Mid-speech Clipping (MSC)*

Сегменти вътре в говорен фрагмент грешно класифицирани като шум. Дефинира се във вида:

$$MSC = \frac{N_M}{N_{SPEECH}}, \tag{4.16}$$

където N_M е брой на говорните сегменти вътре в говорния фрагмент грешно класифицирани като шум и N_{SPEECH} е общ брой говорни сегменти, получени при ръчна детекция.

- *OVER (over hang)*

При преход от говор към шум част от шумовите сегменти грешно се класифицират като говор. Дефинира се във вида:

$$OVER = \frac{N_O}{N_{NON-SPEECH}}, \tag{4.17}$$

където N_O е брой на сегментите грешно класифицирани като говор при преход от говор към шум и $N_{NON-SPEECH}$ е общ брой не-говорни сегменти, получени при ръчна детекция.

- *Noise Detected as Speech (NDS)*

Шум класифициран като говор в границите на шумов фрагмент. Дефинира се във вида:

$$NDS = \frac{N_N}{N_{NON-SPEECH}}, \tag{4.18}$$

където N_N е брой на сегментите с шум класифицирани като говор в границите на шумов фрагмент и $N_{NON-SPEECH}$ е общ брой не-говорни сегменти, получени при ръчна детекция.

- *Correct speech decision made by VAD (CSD)* или *Speech Hit Rate (SHR)*

Вярно класифицирани говорни сегменти. Дефинира се във вида:

$$CSD = \frac{N_{SP}}{N_{SPEECH}}, \tag{4.19}$$

където N_{SP} е броят на коректно класифицираните говорни сегменти и N_{SPEECH} е общ брой говорни сегменти, получени при ръчна детекция.

- *Correct noise decisions made by VAD (CND)* или *Non-speech Hit Rate (NHR)*

Вярно класифицирани не-говорни сегменти. Дефинира се във вида:

$$CND = \frac{N_{NSP}}{N_{NON-SPEECH}}, \tag{4.20}$$

където N_{NSP} е броят на вярно класифицираните не-говорни фрагменти и $N_{NON-SPEECH}$ е общ брой не-говорни сегменти, получени при ръчна детекция.

- *Back-End Clipping (BEC)*

При преход от говор към шум част от говорните сегменти грешно се класифицират като шум. Дефинира се във вида:

$$BEC = \frac{N_S}{N_{SPEECH}}, \tag{4.21}$$

където N_S е броят на говорните сегменти грешно класифицирани като шум при преход от говор към шум; N_{SPEECH} Е общ брой говорни сегменти, получени при ръчна детекция.

- *Front-end adding (FEA)*

Сегменти с шум грешно се класифицират като говор при преход от шум към говор. Дефинира се във вида:

$$FEA = \frac{N_{FEA}}{N_{NON-SPEECH}}, \tag{4.22}$$

където N_{FEA} е броят на сегменти с шум грешно класифицирани като говор при преход от шум към говор и $N_{NON-SPEECH}$ е общ брой не-говорни сегменти, получени при ръчна детекция.

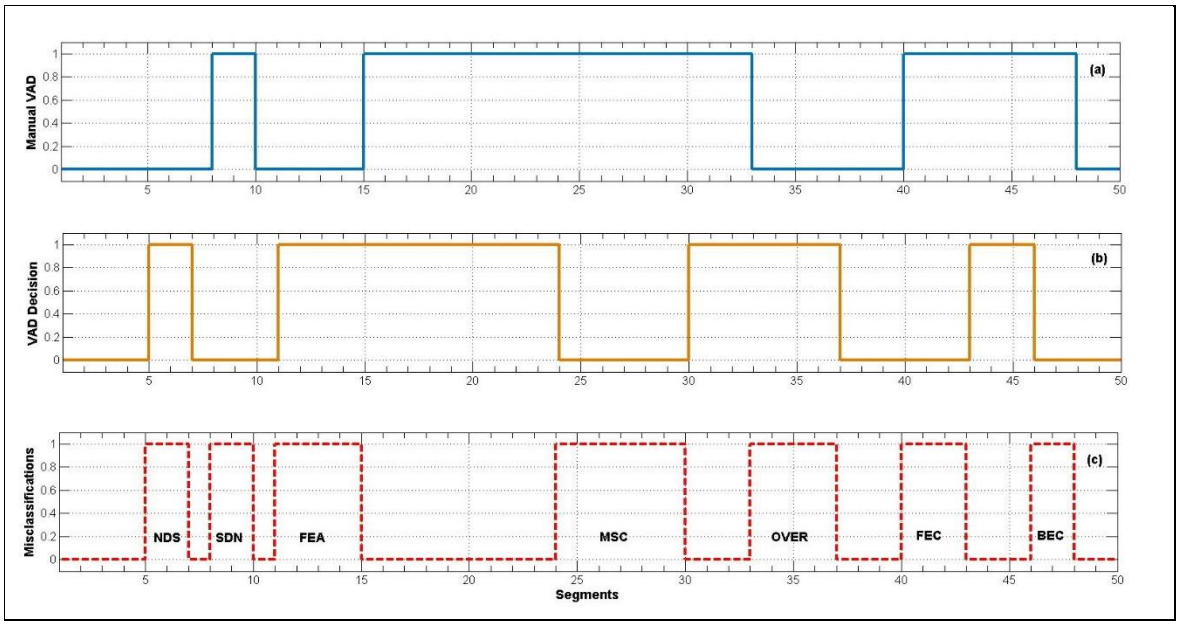
- *Speech Detected as Noise (SDN)*

Говор класифициран като шум в границите на шумов фрагмент. Дефинира се във вида:

$$NDS = \frac{N_{SPN}}{N_{SPEECH}}, \tag{4.23}$$

където N_{SPN} е брой на говорните сегменти класифицирани като шум в границите на шумов фрагмент и N_{SPEECH} е общ брой говорни сегменти, получени при ръчна детекция.

На фиг. 4.1 са показани съответно: (a) примерен ръчно детектиран сигнал, (b) резултат от автоматична детекция на същия сигнал и на (c) –грешна класификация с означени всички изброени по-горе грешки.



Фиг. 4.1. Грешки при детекция: (a) примерен ръчно детектиран сигнал, (b) резултат от автоматична детекция на същия сигнал; (c) грешна класификация с означени всички изброени в текста грешки.

4.4. Оценка на точността при детекция и грешката при разпознаване

В работата са изследвани два детектора на говор, които в действителност са бинарни класификатори. Техните характеристики могат да се анализират обективно чрез ROC (Receiver Operating Characteristics) - графики или/и чрез матрица на грешките (confusion

matrix). В повечето случаи интерпретацията на резултатите от класификация се затруднява значително, ако директно се използват ROC-графики и матрица на грешките. Този факт е довел до въвеждане на скаларни величини, които представят в обобщен вид някои от характеристиките на посочените по-горе два подхода. В работата са използвани подобни величини съответно: при ROC-анализа - F-measure и AUC (Area Under ROC Curve) [Fawcett, 2006], а при матрицата на грешките - ентропията, изчислена на базата на матрицата на грешките - Confusion Entropy (CEN) [Wei et al., 2010].

4.4.1. ROC-анализ

Чрез ROC-графиката (TPR vs FPR) се представят двумерно характеристиките на класификатора. TPR (True Positive Rate) се изчислява като отношение на коректно класифицираните говорни сегменти към всички говорни сегменти. Съответно FPR (False Positive Rate) е отношението на некоректно класифицирани не-говорни сегменти към всички не-говорни сегменти. Площта под ROC-графиката (част от единичен квадрат) е величината AUC и тя има винаги стойност между 0 и 1. При класификация чрез случаен избор и при реални класификатори стойността на AUC не би трябвало да бъде по малка от 0.5. Фактически AUC е глобална мярка за степента на разделимост между разпределянията на оценките (scores) получени за векторите принадлежащи към всеки един от двата класа. Резултатите, получени чрез AUC обаче не са еднозначни. Възможно е при реални условия да се получат ситуации, при които да не съществува съгласуваност между грешката при разпознаване и стойността на AUC. Например - класификатори, при които релациите между грешките и стойностите на AUC са разнопосочни [Fawcett, 2006].

Освен стойността на AUC в работата ще бъде използван и параметъра F-measure (FM). Той е дефиниран като средно хармонична стойност на параметрите *Precision* и *Recall* за положителния клас [Ferri et al., 2009]

$$FM = \frac{2.recall.precision}{recall + precision} \tag{4.24}$$

Където *Recall* е TPR дефинирана по-горе в текста, т.е. точността на детекция на говорните сегменти, докато *Precision* е отношение на вярно класифицираните говорни сегменти към общия брой сегменти класифицирани като говорни или

$$recall = \frac{correctly\ \ classified\ \ \ positives}{total\ \ \ positives} \tag{4.25}$$

$$precsion = \frac{correctly\ \ classified\ \ \ positives}{total\ \ \ prediceted\ \ \ positives} \tag{4.26}$$

Известно е, че при анализ на ефективността на класификатори F-measure е по-добър критерий в сравнение точността на класификация, защото се влияе по-малко от асиметричността на разпределението на данните [Forman et al., 2004].

4.4.2. Матрица на грешките

Известно е, че грешката не е надежден критерий за оценка на качествата на даден алгоритъм за разпознаване. Възможно е например при два различни класификатора да се получи една и съща грешка и то, при положение че използваните данни имат различни разпределения [Delgado et al., 2019]. Чрез грешката е невъзможно да се оценят редица детайли като например как се разграничават данните от различните класове. Подобна информация е полезна в конкретния случай, когато се използва обща невронна мрежа.

За по-детайлен анализ на резултатите от разпознаването в работата е използвана предложената в [Wei et al., 2010] мяра основана на ентропията и изчислена на базата на матрицата на грешките - Confusion Entropy (CEN).

Ако имаме матрица на грешките $\|C\|$ с елемент $c_{i,j}, i=1,...,N; j=1,2,...,N;$, N – брой класове, то вероятността данни от клас i да се класифицират като принадлежащи към клас j се означава като $P_{i.j}^{j}$ или това е относителната честота на данните от клас i които са класифицирани като клас j спрямо всички данни, принадлежащи към клас j. Дефинира се във вида

$$P_{i,j}^{j}=\frac{C_{i,j}}{\sum_{k=1}^{N}\left(C_{j,k}+C_{k,j}\right)-C_{j,j}}, i,j=1,...,N, i\neq j. \tag{4.27}$$

Съответно $P_{i.j}^{i}$ е относителната честота на данните от клас i които са класифицирани като клас j спрямо всички данни, принадлежащи към клас i и се дефинира във вида

$$P_{i,j}^{i}=\frac{C_{i,j}}{\sum_{k=1}^{N}\left(C_{i,k}+C_{k,i}\right)-C_{i,i}}, i,j=1,...,N, i\neq j \tag{4.28}$$

Вероятността за грешка (confusion probability) за клас j е дефинирана във вида

$$P_j=\frac{\sum_{k=1}^{N}\left(C_{j,k}+C_{k,j}\right)-C_{j,j}}{2\sum_{k,l=1}^{N}C_{k,l}-\alpha\sum_{k=1}^{N}C_{k,k}};\quad \alpha=\begin{cases}0.5; \text{ if } N=2\\ 1; \text{ if } N>2\end{cases} \tag{4.29}$$

Ентропията на грешката (confusion entropy) асоциирана към клас j се дефинирана като

$$CEN_j = -\sum_{k=1,k\neq j}^{N}\left(P_{j,k}^{j}\,log_{2(N-1)}(P_{j,k}^{j}) + P_{k,j}^{j}\,log_{2(N-1)}(P_{k,j}^{j})\right). \quad (4.30)$$

Общата (за всички класове) ентропия на грешката (overall confusion entropy) е

$$CEN = \sum_{j=1}^{N} P_j.CEN_j\,. \quad (4.31)$$

Така определената мяра (т.е. confusion entropy) има две особености. Първата е, че при нейното определяне се вземат предвид само грешно класифицираните данни. Това означава, че малък брой на грешни класификации ще доведе до малки стойности на ентропията. Ако всички данни са коректно класифицирани то тази мяра има стойност нула, т.е. малка ентропия съответства на висока точност. Ако няма вярно класифицирани данни то ентропията става равна на 1. Освен това при различни тестове е възможно извън диагоналните елементи да варират в широки граници докато сумата по диагонала да остава една съща (т.е. еднаква точност). Или при класификация ентропията има по-силно изразени дискриминиращи свойства, отколкото точността. Освен това чрез ентропията, асоциирана за всеки клас може да прецени доколко има застъпване на данните между класовете [Wei et al., 2010], [Delgado et al., 2019].

4.5. Идентификация на диктори независимо от текста

Въпроса за избор на подход за независимо от текста разпознаване на диктори е дискутиран подробно в редица публикации [Togneri et al., 2011], [Hansen et al., 2015]. В конкретния случай, ако се използват генеративни алгоритми от вида GMM или HMM то ограничения обем на използваните данни и силната им зашуменост (записи от телефонен канал) биха довели до затруднения при формиране на моделите (вероятностните разпределяния) при отделните диктори [Fazel et al., 2011]. По-подходящи в случая са дискриминативни алгоритми като NN или SVM при които за дадени данни в режим на обучение се минимизира грешката при класификация и се моделира само границата между класовете.

В работата е прието като класификатор да се използва многослоен перцептрон с един скрит слой и алгоритъм за обучение чрез обратно разпространение на грешката. Мотивация за подобно решение са аргументите представени по-горе в текста както и резултатите от изследванията на автора описани в [Ouzounov, 2009]. Освен това трябва да се има предвид, че в изследването архитектурата и алгоритъма за обучение на многослойния перцептрон са пряко свързани с ограничения обем на говорни данни. В работата са използвани телефонни записи от корпуса BG-SRDat [виж гл. 5] и [Ouzounov,

2003] които са с обща продължителност от около 4 минути за всеки един от дикторите (в независим от текста режим).

В работата акцентът е върху модула за детекция на говор. По подобие на подхода в гл. 3 и тук за всеки от предложените параметри се реализира отделен детектор на говор, който участва самостоятелно в системата за разпознаване на диктори и по този начин се анализира неговото влияние върху грешката от идентификация.

4.6. Детектор на говор – VAD-1

Предложения в работата VAD алгоритъм включва 3 етапа – определяне на параметри, класификация чрез многослоен перцептрон и прагов алгоритъм (thresholding scheme).

4.6.1. Използвани признаци

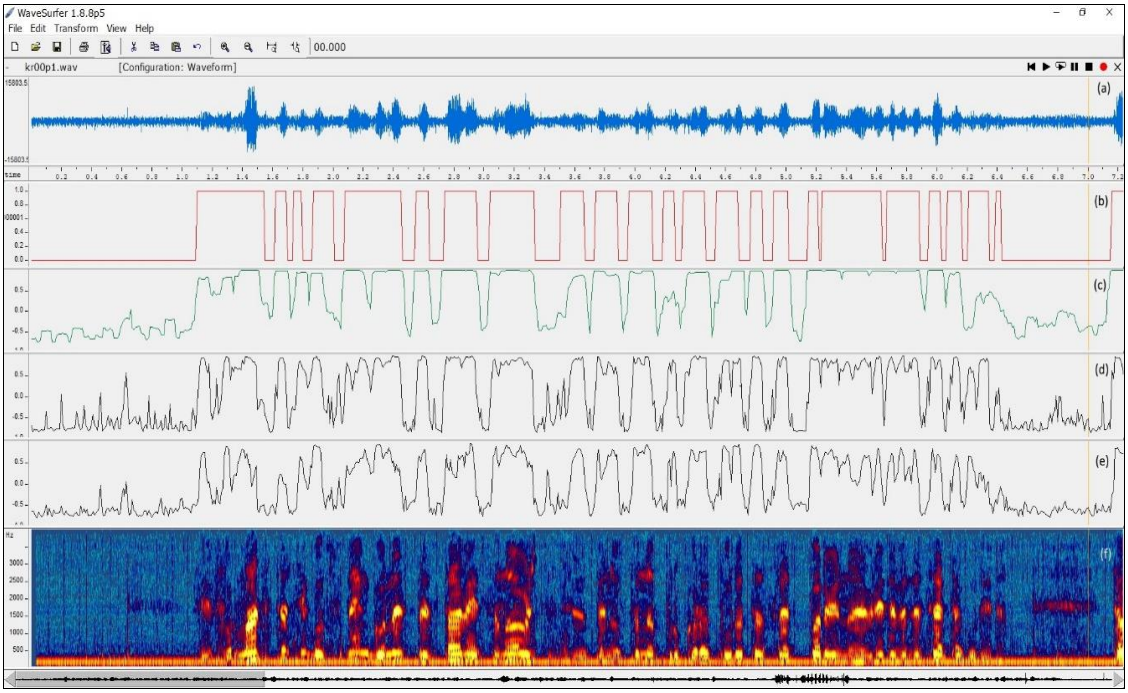
В детектора на говор са използвани четири референтни признака описани в т. 2 и два предложени от автора и описани в т. 2.1.4.2 и т. 2.1.4.3. За всеки признак е реализиран отделен детектор. Признаците са:

- базов среден-делта признак (Basic Mean-Delta - BMD feature) – т. 2.1.4.2 (предложен от автора);
- модифициран среден-делта признак (Modified Mean-Delta - MMD feature) – т. 2.1.4.3 (предложен от автора);
- многолентова спектрална ентропия – (Multi-band Spectral entropy - MBSE) – т. 4.2.2, (референтен);
- честотно филтриран признак – (Frequency Filtered – FF feature) – т. 4.2.3, (референтен);
- относителна спектрална разлика – (Relative Spectral Difference - RSD) – т. 4.2.3, (референтен);
- МЕЛ-кепстър с индексен лифтер – (Index Weighted Mel-Frequency Cepstral Coefficients – IW-MFCC) – т. 4.2.4, (референтен).

За всички признаци говорния сигнал е разделен на сегменти с дължина 30 ms и изместване 10 ms и се умножава с прозорец на Хеминг. При определяне на спектъра е използвано FFT с дължина 512 точки. Размер на векторите за всички параметрични представяния е 14.

На фиг. 4.2 са показани - начален фрагмент с дължина около 7 sec от телефонен запис и стойностите във времето получени на изходните неврони за съответните VAD детектори както следва: (a) говорен сигнал, (b) ръчна сегментация, (c) VAD с признак BMD, (d) VAD с МЕЛ-кепстър с индексен лифтер, (e) VAD с признак RSD и (f)

спектрограма. Тези контури са получени за тестов файл, анализиран с вече обучения многослоен перцептрон. При генериране на графиката са използвани WaveSurfer [WaveSurfer, online] и Paint.Net [Paint.net, online].



Фиг. 4.2. Фрагмент от телефонен запис от корпуса BG-SRDat и контури получени на изходните неврони на съответните VAD детектори: (a) говорен сигнал, (b) ръчна сегментация, (c) VAD с признак BMD, (d) VAD с МЕЛ-кепстър с индексен лифтер, (e) VAD с признак RSD и (f) спектрограма.

4.6.2. Многослоен перцептрон

Структурата на използвания в детектора МП е от вида 14-20-1. Невронната мрежа има един скрит слой с 20 неврона и изходен слой с един неврон. Функцията на активност (activation function) за всички неврони е хиперболичен тангенс. Обучението е реализирано чрез алгоритъма RProp в пакетен режим (batch mode) и със стойности на параметрите избрани съгласно препоръките в [Demuth et al., 2009] и [LeCun et al., 2012]. Входния вектор е с размерност 14, при обучение са използвани референтни нива за изходните неврони (target levels) [-0.9; 0.9] съответно за шум/говор. При обучение се установяват следните параметри: епохи (epochs) 500; крайна грешка при обучение (performance goal) 1.0e-5; прекъсване на обучението, когато грешката от валидация е увеличена за повече от 10 итерации [Demuth et al., 2009].

4.6.3. Прагов алгоритъм

При детекция бинарното решение се получава чрез праг, с който се сравнява стойността, получена в изходния неврон. Прието е прага да бъде определен чрез алгоритъма на Otsu [Kisku et al., 2014] като изчисляването му се извършва поотделно за всеки тестов файл.

Предварителните експерименти показаха, че стойностите в изходния неврон, получени в границите на даден тестов файл формират хистограма с форма близка до бимодално разпределение. В този случай прага на Otsu се оказва подходящ за разделяне на двата класа (говор/шум). Също така бяха реализирани експерименти, при които се използваха съответно фиксирания и адаптивния праг, предложени от автора в гл. 3. В конкретния случай получените резултати не бяха по-добри от тези получени с прага на Otsu и затова те не са включени в изложението.

4.6.4. Говорни данни, използвани при VAD-1

Данните, преназначени за детектора на говор са избрани от корпуса BG-SRDat и са разделени в три групи – за обучение, валидация и тестване. Първата група включва 24 файла, а втората 12 файла. Или за обучение се използват около 70000 сегмента с говор и около 40000 с не-говор. Съответните групи сегменти, предназначени за валидация са около два пъти по-малко. Като еталонни последователности са използвани ръчно маркирани данни. Получените маркери (0/1) на тестовите данни впоследствие се подават като еталонни последователности (VAD targets) при обучение на невронната мрежа за разпознаване на диктори.

4.6.5. Определяне на точността при детекция

За всички използвани параметри в таблица 4.1 са показани стойностите в проценти на грешките при детекция (виж т. 4.3). Те са изчислени като претеглени средни стойности на резултатите от всички тествани файлове – общо 270 файла.

В последните четири реда на таблица 4.1 са показани съответно точностите на детекция на говорните (SHR) и на не-говорните сегменти (NHR) и стойностите на параметрите F-measure и AUC. Оказва се, че няма параметър за детекция на говор, при който точностите на детекция и двата критерия за оценка да имат едновременно максимални стойности. При два от параметрите MMD и IW-MCFF се наблюдават по две максимални стойности: съответно при MMD - SHR и F-measure, а при IW-MCFF – NHR и AUC. Не е възможно само чрез стойностите на SHR, NHR, F-measure и AUC да се оцени предварително как точността на детекция ще влияе върху точността при разпознаване.

Таблица 4.1. Корпус BG-SRDat – VAD-грешки и VAD-точност в проценти и стойности на F-measure и AUC

		Features					
No.	**Errors**	**BMD**	**MMD**	**IW-MFCC**	**RSD**	**FF**	**MBSE**
1	**NDS**	**3.0460**	3.0811	5.3943	5.5603	5.4806	7.0924
2	**SDN**	1.2672	1.3016	**0.1380**	0.2218	0.2188	0.4281
3	**FEA**	7.3957	7.6403	3.2107	3.4770	3.3037	**2.9677**
4	**MSC**	4.9555	**4.6617**	8.5554	8.0090	7.8394	11.6752
5	**OVER**	4.1887	4.2373	2.5303	2.5641	**2.2968**	2.7926
6	**FEC**	2.8267	**2.6551**	3.1080	3.1519	3.2129	4.4374
7	**BEC**	4.7662	**4.5610**	4.8369	5.2852	5.3564	6.1939
	Accuracy	**BMD**	**MMD**	**IW-MFCC**	**RSD**	**FF**	**MBSE**
1	**SHR**	86.0038	**86.6680**	83.3382	83.3661	83.3671	77.2084
2	**NHR**	81.3985	80.9349	**88.0417**	87.0674	87.7306	85.6802
3	**F-Measure**	0.8753	**0.8780**	0.8768	0.8745	0.8762	0.8334
4	**AUC**	0.9028	0.9043	**0.9245**	0.9183	0.9221	0.8701

Освен това характера на грешките при детекция поместени в първите 7 реда влияят различно при формиране на модела на диктора. Например стойността на NDS е показател доколко VAD алгоритъма грешно класифицира шума като говор – или стойността на тази грешка показва колко шумови фрагмента грешно се използват в модела на диктора. Чрез OVER се оценява доколко алгоритъма се справя с детекцията на фонеми с ниско ниво в края на произнасянето. Подобни фонеми обикновено грешно се класифицират като шум и се изрязват от края на фразата. FEC служи за оценка на това доколко алгоритъма се справя с рязкото увеличаване на нивото в началото на произнасянето. В този случай често се изрязват начални сегменти. Обикновено FEC и MSC имат значителни стойности, когато SNR е много ниско.

4.7. Система за идентификация на диктори при VAD-1

Използваната в работата система за независима от текста идентификация на диктори включва 3 модула – предварителна обработка, класификация чрез многослоен перцептрон и модул за вземане на решение чрез анализ на последователност от супра-сегменти (supra-segments decision scheme). Блоковата схема на системата за разпознаване на диктори включваща детайли за VAD-модула е показана на фиг. 4.3.

4.7.1. Предварителна обработка

Модула за предварителната обработка включва два под-модула – за детекция на говор (описан в т. 4.6) и за определяне на кепстрални параметри.

Като кепстрални параметри са използвани 14 коефициента (без нулевия) на МЕЛ-кепстъра получени чрез 24-лентови филтри с еднаква площ. Обработват се само сегментите, за които VAD-детектора е приел решение, че съдържат говор. Формирането на последователността от вектори, която се подава за обучение, валидация и тестване на

НМ включва следните обработки – сегментиране на времевата последователност (30 ms/10 ms), синхронизация с VAD-модула и маркиране на говорните сегменти - и само за говорните сегменти - умножение с прозорец на Хеминг, определяне на МЕЛ-кепстъра, центриране на кепстралната последователност (CMS) в границите на отделен файл и прилагане на индексен лифтер върху така получените кепстрални коефициенти.

4.7.2. Многослоен перцептрон

Както бе посочено по горе в текста в работата е прието като класификатор да се използва многослоен перцептрон с един скрит слой и алгоритъм за обучение чрез обратно разпространение на грешката.

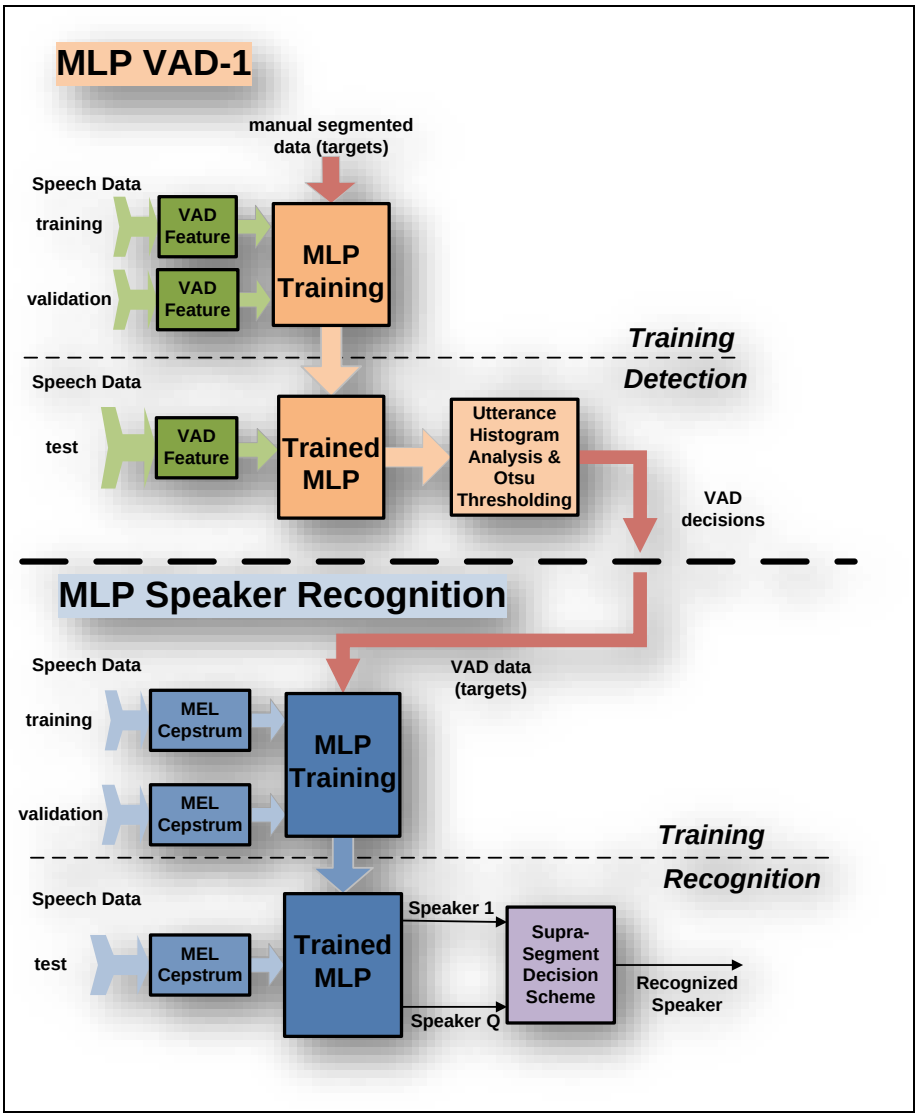
Броят на дикторите, които ще бъдат разпознавани е 12 и структурата на използвания в работата МП е от вида 14-120-12. Размера на входния вектор е 14, броят на невроните в скрития слой 120 и броят на изходните неврони – 12. Функцията на активност (activation function) за всички неврони е хиперболичен тангенс. Обучението е реализирано чрез алгоритъма RProp в пакетен режим (batch mode) и стойностите на параметрите са избрани съгласно препоръките в [Demuth et al., 2009]. При обучение са използвани референтни нива за изходните неврони (target levels) [-0.76; 0.76]. При обучение се установяват следните параметри: епохи (epochs) 500; крайна грешка при обучение (performance goal) 1.0e-5; обучението спира, ако се попадне в глобален минимум на средно квадратичната грешка или когато грешката от валидация се увеличава за повече от 10 итерации [Demuth et al., 2009].

За се компенсира влиянието на случайната инициализация в многослойния перцептрон, тук е приложен алгоритъма на многократните стартирания (multiple runs scheme) и е приета схема 5x10 [Kuncheva, 2014]. В работата детектора на говор се стартира 5 пъти като при всяко стартиране (детекция) се извършва и 10 пъти стартиране (с данни от текущата детекция) на МСП за обучение и идентификация на диктори. Процедурата се повтаря за всеки от изследваните параметри.

4.7.3. Говорни данни използвани при разпознаване на диктори

Данните предназначени за разпознаване на диктори са избрани от корпуса BG-SRDat и са разделени в три групи - за обучение, валидация и тестване. При формиране на данните за обучение са използвани еднакъв брой говорни сегменти за всеки клас. И тъй като файловете са с различен текст и респективно с различна дължина то е прието броя на говорните сегменти от един файл да бъде 1300. За всеки диктор са използвани 2 файла или обучаваща последователност за него ще съдържа 2600 говорни сегмента, получени чрез случаен избор от всички говорни сегмента съдържащи се в двата файла. Избора на

данни за валидация е по аналогичен начин, само че се използват данни от един файл за диктор (т.е. 1300 сегмента).



Фиг. 4.3. Блоковата схема на системата за разпознаване на диктори включваща детайли за VAD-1 модула.

4.7.4. Модул за вземане на решение

Разпознаването на диктори е реализирано чрез анализ на супра-сегменти. Дължината на един супра-сегмент е 200 сегмента (2 секунди). Той се измества без застъпване по дължината на тествания файл (анализират се само говорни сегменти). Идентификацията на диктора се осъществява за всеки супра-сегмент поотделно. Анализира се вектора, получен чрез усредняване на всички изходни вектори на перцептрона за дадения супра-сегмент. Разпознатия клас е този при който е получена максималната стойност във вектора.

4.7.5. Експериментални резултати

Резултатите от идентификацията на диктори са представени като матрица на грешките (confusion matrix). По-долу в текста за всеки един от признаците са показани отделни таблици с матрица на грешките, както следва – признаци BMD, MMD, IW-MFCC, RSD, FF и MBSE и таблици 4.2-4.7.

Таблица 4.2. Матрица на грешките при идентификация на диктори за признак BMD

№	True class	Recognized class												Errors	Tests
		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12		
1	**Spk #1**	**614**	8	2	0	0	0	1	0	2	0	0	3	16	630
2	**Spk #2**	54	**629**	7	0	0	0	0	0	18	0	0	32	111	740
3	**Spk #3**	0	0	**730**	0	0	0	0	0	0	0	0	0	0	730
4	**Spk #4**	41	9	9	**474**	171	13	0	14	2	6	0	1	266	740
5	**Spk #5**	2	0	0	0	**424**	2	6	0	0	1	5	0	16	440
6	**Spk #6**	7	1	0	34	73	**486**	1	0	2	0	6	0	124	610
7	**Spk #7**	0	0	1	21	14	48	**596**	6	1	8	1	4	104	700
8	**Spk #8**	1	4	0	0	0	3	1	**603**	0	108	0	0	117	720
9	**Spk #9**	0	3	6	37	49	7	3	1	**516**	0	1	7	114	630
10	**Spk #10**	2	0	3	0	4	27	98	18	0	**411**	21	6	179	590
11	**Spk #11**	0	0	1	0	4	77	2	0	1	0	**595**	0	85	680
12	**Spk #12**	0	0	0	0	0	0	0	2	0	0	0	**628**	2	630
13	**Errors**	107	25	29	92	315	177	112	41	26	123	34	53	**1134**	**7840**
14	**Tests**	721	654	759	566	739	663	708	644	542	534	629	681	**7840**	**14.46**

Таблица 4.3. Матрица на грешките при идентификация на диктори за признак WMD

№	True class	Recognized class												Errors	Tests
		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12		
1	**Spk #1**	**620**	17	2	0	0	0	0	0	0	0	0	1	20	640
2	**Spk #2**	63	**645**	5	0	0	0	0	1	6	0	2	28	105	750
3	**Spk #3**	0	0	**730**	0	0	0	0	0	0	0	0	0	0	730
4	**Spk #4**	46	5	13	**418**	221	10	0	12	2	2	0	1	312	730
5	**Spk #5**	16	0	0	0	**394**	18	6	0	0	0	6	0	46	440
6	**Spk #6**	4	0	0	37	53	**500**	5	0	1	0	20	0	120	620
7	**Spk #7**	0	0	0	17	16	42	**631**	6	1	6	0	1	89	720
8	**Spk #8**	0	7	0	0	0	4	0	**600**	0	109	0	0	120	720
9	**Spk #9**	2	0	3	50	56	15	3	9	**502**	0	5	5	148	650
10	**Spk #10**	6	2	1	0	10	54	81	24	0	**409**	11	2	191	600
11	**Spk #11**	1	0	3	0	5	100	4	0	0	0	**577**	0	113	690
12	**Spk #12**	0	0	0	0	0	0	0	1	0	1	0	**648**	2	650
13	**Errors**	138	31	27	104	361	243	99	53	10	118	44	38	**1266**	**7940**
14	**Tests**	758	676	757	522	755	743	730	653	512	427	621	686	**7940**	**15.94**

Таблица 4.4. Матрица на грешките при идентификация на диктори за признак IW-MFCC

№	True class	Recognized class												Errors	Tests
		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12		
1	**Spk #1**	**530**	8	2	0	0	0	0	0	0	0	0	0	10	540
2	**Spk #2**	16	**575**	19	0	0	0	0	0	19	0	2	19	75	650
3	**Spk #3**	0	0	**750**	0	0	0	0	0	0	0	0	0	0	750
4	**Spk #4**	37	23	46	**385**	154	4	4	40	2	9	0	6	325	710
5	**Spk #5**	1	0	0	3	**348**	9	29	0	0	0	0	0	42	390
6	**Spk #6**	30	3	0	39	38	**370**	6	6	0	0	28	0	150	520
7	**Spk #7**	0	0	0	14	10	5	**611**	8	0	2	0	0	39	650
8	**Spk #8**	1	7	0	6	1	37	0	**538**	0	146	4	0	202	740
9	**Spk #9**	0	2	7	34	39	6	38	7	**379**	1	0	7	141	520
10	**Spk #10**	0	0	0	5	23	194	67	18	0	**169**	24	0	331	500
11	**Spk #11**	0	0	0	0	1	46	5	0	0	0	**598**	0	52	650
12	**Spk #12**	0	0	0	0	1	0	0	5	0	1	0	**653**	7	660
13	**Errors**	85	43	74	101	267	301	149	84	21	159	58	32	**1374**	**7280**
14	**Tests**	615	618	824	486	615	671	760	622	400	328	656	685	**7280**	**18.87**

Таблица 4.5. Матрица на грешките при идентификация на диктори за признак RSD

№	True class	Recognized class												Errors	Tests
		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12		
1	**Spk #1**	**485**	8	9	0	0	0	0	0	0	0	0	8	25	510
2	**Spk #2**	19	**590**	22	2	0	0	0	1	16	0	0	50	110	700
3	**Spk #3**	0	0	**760**	0	0	0	0	0	0	0	0	0	0	760
4	**Spk #4**	55	9	37	**460**	50	3	0	20	3	8	0	5	190	650
5	**Spk #5**	4	0	0	1	**327**	14	19	0	0	0	5	0	43	370
6	**Spk #6**	5	2	0	36	30	**393**	4	9	0	1	50		137	530
7	**Spk #7**	0	0	0	7	12	4	**623**	2	1	1	0	0	27	650
8	**Spk #8**	1	11	0	2	2	42	5	**522**	0	171	4	0	238	760
9	**Spk #9**	0	0	7	49	46	5	45	9	**277**	6	4	2	173	450
10	**Spk #10**	2	0	0	4	17	266	114	8	0	**80**	9	0	420	500
11	**Spk #11**	4	6	0	0	0	54	3	0	1	0	**620**	2	70	690
12	**Spk #12**	0	0	0	0	0	0	0	2	0	0	0	**738**	2	740
13	**Errors**	90	36	75	101	157	388	190	51	21	187	72	67	**1435**	**7310**
14	**Tests**	575	626	835	561	484	781	813	573	298	267	692	805	**7310**	**19.63**

Таблица 4.6. Матрица на грешките при идентификация на диктори за признак FF

№	True class	Recognized class												Errors	Tests
		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12		
1	**Spk #1**	**449**	32	11	0	0	0	0	0	0	0	1	7	51	500
2	**Spk #2**	25	**566**	26	0	3	0	3	2	16	0	1	48	124	690
3	**Spk #3**	0	0	**760**	0	0	0	0	0	0	0	0	0	0	760
4	**Spk #4**	147	7	33	**388**	83	2	0	25	3	11	0	11	322	710
5	**Spk #5**	5	0	0	8	**299**	15	21	0	0	0	2	0	51	350
6	**Spk #6**	3	4	0	79	12	**339**	6	4	0	1	62	0	171	510
7	**Spk #7**	0	0	0	23	6	8	**579**	4	0	0	0	0	41	620
8	**Spk #8**	8	13	1	0	1	32	0	**578**	0	158	9	0	222	800
9	**Spk #9**	0	2	9	78	53	7	35	9	**295**	7	1	4	205	500
10	**Spk #10**	1	0	0	4	8	189	79	8	0	**92**	9	0	298	390
11	**Spk #11**	0	1	1	0	0	32	2	0	0	0	**524**	0	36	560
12	**Spk #12**	1	0	0	0	0	0	0	1	0	1	0	**587**	3	590
13	**Errors**	190	59	81	192	166	285	146	53	19	178	85	70	**1524**	**6980**
14	**Tests**	639	625	841	580	465	624	725	631	314	270	609	657	**6980**	**21.83**

Таблица 4.7. Матрица на грешките при идентификация на диктори за признак MBSE

№	True class	Recognized class												Errors	Tests
		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12		
1	**Spk #1**	**399**	27	4	0	0	0	0	0	0	0	0	0	31	430
2	**Spk #2**	32	**529**	45	0	2	0	0	2	21	0	1	18	121	650
3	**Spk #3**	0	0	**770**	0	0	0	0	0	0	0	0	0	0	770
4	**Spk #4**	176	3	38	**343**	111	0	0	13	6	20	0	0	367	710
5	**Spk #5**	1	0	0	4	**252**	17	56	0	0	0	0	0	78	330
6	**Spk #6**	2	23	0	17	1	**299**	23	6	0	0	29	0	101	400
7	**Spk #7**	0	0	0	0	0	0	**530**	0	0	0	0	0	0	530
8	**Spk #8**	2	26	1	3	16	108	1	**450**	0	138	24	1	320	770
9	**Spk #9**	0	17	2	25	44	3	55	27	**187**	6	14	0	193	380
10	**Spk #10**	0	0	0	2	8	228	174	0	0	**48**	0	0	412	460
11	**Spk #11**	6	0	12	0	0	28	0	0	0	1	**583**	0	47	630
12	**Spk #12**	1	0	0	1	0	0	0	46	0	0	0	**712**	48	760
13	**Errors**	220	96	102	52	182	384	309	94	27	165	68	19	**1718**	**6820**
14	**Tests**	619	625	872	395	434	683	839	544	214	213	651	731	**6820**	**25.19**

За всеки признак в таблица 4.8 са показани стойностите на ентропията за отделния клас, както и общата ентропия. Грешката при разпознаване е включена в последния ред на таблицата.

Таблица 4.8. Общата ентропия и ентропията за всеки един диктор

		Features					
No.	**CEN**	**BMD**	**MMD**	**IW-MFCC**	**RSD**	**FF**	**MBSE**
1	**Sp #1**	0.1149	0.1378	**0.1041**	0.1353	0.1997	0.1902
2	**Sp #2**	0.1270	**0.1229**	0.1333	0.1487	0.1835	0.2165
3	**Sp #3**	0.0354	**0.0321**	0.0611	0.0645	0.0710	0.0805
4	**Sp #4**	**0.2719**	0.2903	0.3649	0.2841	0.3849	0.3276
5	**Sp #5**	**0.2437**	0.2874	0.2779	0.2615	0.2724	0.2944
6	**Sp #6**	**0.2622**	0.2962	0.3577	0.3444	0.3613	0.3579
7	**Sp #7**	0.1795	**0.1592**	0.1687	0.1652	0.1667	0.1928
8	**Sp #8**	**0.1244**	0.1373	0.2195	0.2139	0.2008	0.3118
9	**Sp #9**	**0.1557**	0.1663	0.2089	0.2763	0.2802	0.3668
10	**Sp #10**	**0.2628**	0.2690	0.4029	0.4254	0.4195	0.4100
11	**Sp #11**	0.1062	0.1340	**0.1061**	0.1301	0.1233	0.1207
12	**Sp #12**	0.0615	**0.0448**	0.0463	0.0590	0.0749	0.0567
13	**Overall CEN**	**0.1582**	0.1686	0.1917	0.1913	0.2124	0.2191
14	**Recog.Err.[%]**	**14.46**	15.94	18.87	19.63	21.83	25.19

В таблица 4.8 се наблюдава съгласуваност между грешката и общата ентропия за първите два и последните два признака. При третия и четвъртия признак (IW-MFCC и RSD) не е налице такава съгласуваност. В случая важен е фактът, че минималната грешка от разпознаване и минималната обща CEN при BMD и MMD са частично в синхрон с резултатите получени при детекция на говор, а именно максималните стойности на F-measure и SHR наблюдавани в таблица 4.1 за параметъра MMD. Интересно е различието между резултатите на BMD и MMD, което се обяснява с това, че MMD е получена чрез допълнителна филтрация (smoothing) на BMD (виж т. 2.1.4.3).

4.8. Детектор на говор – VAD-2

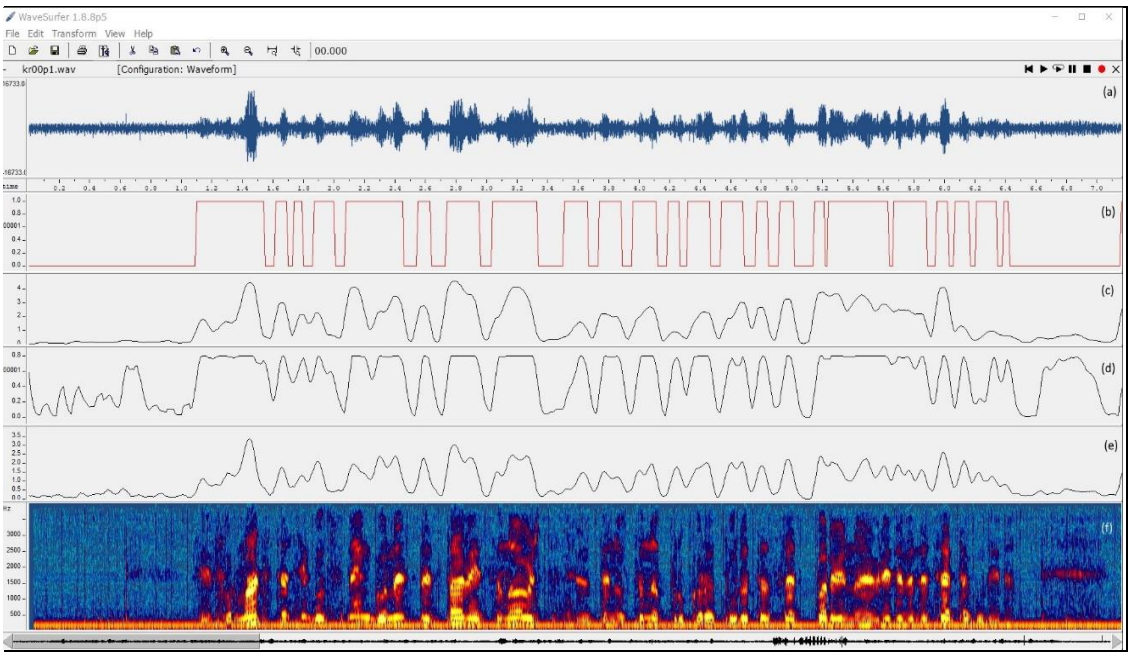
Предложения в работата алгоритъм VAD-2 включва 2 етапа – определяне на признаци и прагов алгоритъм (thresholding scheme).

4.8.1. Изчисляване на признаци

Тук са използвани три референтни признака и един предложен от автора. За всеки признак е реализиран отделен детектор. Признаците са:

- log-GDMD – описан в т.2.2.4 (предложен от автора);
- статистическа оценка $\Gamma(m)$ - ф-ла 1.25 - получена чрез алгоритъма на Sohn (описан в т. 1.2.3.1.1) – (референтен);
- признак SAE – получен чрез алгоритъма на Wu (описан в т. 4.2.1) – (референтен);
- признак LTSD – получен чрез алгоритъма на Ramirez (описан в т. 3.2.4) – (референтен).

За всички признаци говорния сигнал е разделен на сегменти с дължина 30 ms и изместване 10 ms и се умножава с прозорец на Хеминг. При определяне на спектъра е използвано FFT с дължина 512 точки. На фиг. 4.4 са показани - начален фрагмент с дължина около 7 sec от телефонен запис и контури на използваните параметри както следва: (a) говорен сигнал, (b) ръчна сегментация, (c) log-GDMD, (d) контур от алгоритъма на Sohn, (e) контур на SAE – алгоритъм на Wu и (f) спектрограма. Използвания телефонен запис е същият както този на фиг. 4.2. При генериране на графиката са използвани WaveSurfer [WaveSurfer, online] и Paint.Net [Paint.net, online].



Фиг.4.4. Фрагмент от телефонен запис от корпуса BG-SRDat и контури получени за използваните параметри: (a) говорен сигнал, (b) ръчна сегментация, (c) log-GDMD, (d) алгоритъм на Sohn, (e) алгоритъм на Wu и (f) спектрограма.

4.8.2. Прагов алгоритъм

При детекция бинарното решение се получава чрез праг, с който се сравнява стойността на контура. В работата е прието да се използват прагове, получени чрез алгоритмите, предложени от автора в т. 3.4.1 и т. 3.4.2 – съответно фиксирани и адаптивни прагови стойности. В действителност фиксирани прагове се прилагат при параметрите log-GDMD, Sohn и SAE. Алгоритъма LTSD включва собствен праг, който се адаптира спрямо нивото на шума. Адаптивния праг от т. 3.4.2 е използван само за параметъра log-GDMD (отбелязано е в съответната таблица). Необходимо е да се уточни, че в алгоритмите т. 3.4.1 и т. 3.4.2 се изчисляват два прага (първи и втори праг или нисък и висок праг), но тук при всички експерименти е използван само един от тях и това е първият праг.

4.8.3. Говорни данни, използвани при VAD-2

Говорните данни, използвани при анализ на точността на VAD-2 са различни от тези при VAD-1. Основната причина е, че VAD-1 е реализиран като класификатор, който изисква данни за обучение, валидация и тестване – именно затова при него са използвани данни само от BG-SRDat. При VAD-2 който е с прагова логика е прието при оценка на точността освен от BG-SRDat да бъдат използвани и говорни данни на английски език избрани от корпусите – на Dan Ellis [Dan Ellis, online] и NOIZEUS [NOIZEUS, online]. Недостатъчния обем от данни в тези корпуси обаче не позволява те да бъдат използвани за обучение на МСП във VAD-1. Данните, преназначени за детектора на говор избрани от корпуса BG-SRDat съдържат общо 54 файла от 12 диктора. Записите са с различна продължителност като тя е в границите между 30 и 80 секунди. Данните избрани от корпуса NOIZEUS са 240 файла – 6 диктора x 5 кратки фрази (около 3 секунди) x 8 вида шум добавен допълнително към оригиналния чист сигнал. Избрани са данни при които SNR=5 dB. Данните избрани от корпуса на Dan Ellis [Dan Ellis, online] включват цифри от TIDIGITS и изречения от TIMIT. Оригиналните данни са изкуствено зашумени. Общия брой на файловете е 168. Всички данни на английски език са ръчно сегментирани.

4.8.4. Определяне на точност при детекция.

Резултатите от детекция са представени в четири таблици – първите две таблица – 4.9 и 4.10 - съдържат резултати, получени с данни от BG-SRDat, таблица 4.11 включва резултати, получени с данни от NOIZEUS и таблица 4.12 – с данни от корпуса на Dan Ellis. В таблица 4.9 са показани стойностите в проценти на точността при детекция SHR и NHR, съответно за говорни и шумови сегменти и съответно стойността на AUC. В таблица 4.10 са показани основните грешки при детекция (също в проценти) – FEC, BEC, MSC, SDN, FEA, OVER и NDS (виж т. 4.3). Всички стойности са изчислени като

претеглени средни стойности на резултатите от всички файлове. При определяне на фиксирания праг (fixed2thr) във ф-ла 3.20 са използвани стойности на α, съответно за GDMD и Wu - 0, 0.1, 0.2, 0.3, 0.4 и 0.5, и за алгоритъма на Sohn - 0.2, 0.3, 0.4, 0.5, 0.6 и 0.7. Избора на диапазоните на изменение на α се основава на резултатите получени при разпознаване (виж таблица 4.13). Минималната грешка при идентификация се получава именно когато α се изменя в посочените граници. С цел да се избегне претоварване на изложението с излишни данни в последните две таблици са включени само стойностите на AUC.

Таблица 4.9. Корпус BG-SRDat – точността на детекция в проценти и стойности на AUC

No.	Features	Accuracy	fixed1thr					
			0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
1	**GDMD**	**SHR**	89.8828	84.4044	78.5717	**72.6076**	66.4045	59.9798
		NHR	75.6189	80.9857	84.7817	**87.7255**	90.2400	92.2755
		AUC	0.9143					
2	**Wu**	**SHR**	94.6902	91.2574	**87.2644**	82.4564	77.0072	70.9068
		NHR	72.8936	80.9327	**86.0908**	89.8636	92.4576	94.3948
		AUC	0.8953					
3	**GDMD adapt1thr**	**SHR**	88.697558					
		NHR	76.970031					
		AUC	0.9143					
4	**LTSD**	**SHR**	**99.9975**					
		NHR	11.2401					
		AUC	0.8200					
5	**Sohn**	**fixed1thr**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**
		SHR	97.0310	96.1395	94.8981	93.5248	**91.5351**	89.2796
		NHR	61.1788	65.7420	70.2467	74.0062	**77.9767**	81.4161
		AUC	0.9116					

Представени резултати в таблици 4.9 и 4.10 трудно биха могли да бъдат интерпретирани еднозначно. Видно е, че съществува връзка (заради праговете) между точността, с която се определят говорните сегменти и точността при шумовите сегменти. Например чрез алгоритъма LTSD се локализират почти без грешка говорните сегменти (точност 99.99%). Но това става за сметка на много ниската точност при детекция на фрагментите с шум (11.24%). В този случай при обучение на модела на диктора ще бъдат добавени значително количество шумови сегменти интерпретирани като говор.

Резултатите, получени за корпуса NOIZEUS (таблица 4.11) показват, че за четири от осемте вида шум при признака log-GDMD се получава максимална стойност на AUC. При останалите четири вида шум доминира LTSD признака. Необходимо е да се уточни, че само при данните от NOIZEUS е увеличена дължината на LTSE филтъра при изчисляване log-GDMD във ф-ла (2.50). Увеличената дължина на филтъра изглажда допълнително получения контур и компенсира по този начин локалните изменения на

нивото на шума. При данните от другите два корпуса (BG-SRDat и DanEllis) подобно увеличаване на дължината на филтъра води до влошаване на резултатите.

Съгласно резултатите показани в таблица 4.12 за данни от корпуса на Dan Ellis стойността на AUC е максимална при признака log-GDMD.

На фиг. 4.5 са показани средните DET графики [Beigi, 2011] визуализиращи изменението на грешката при детекция на говор за данните от корпуса на Dan Ellis и изследваните тук признаци.

Таблица 4.10. Корпус BG-SRDat - средни стойности на грешките при детекция

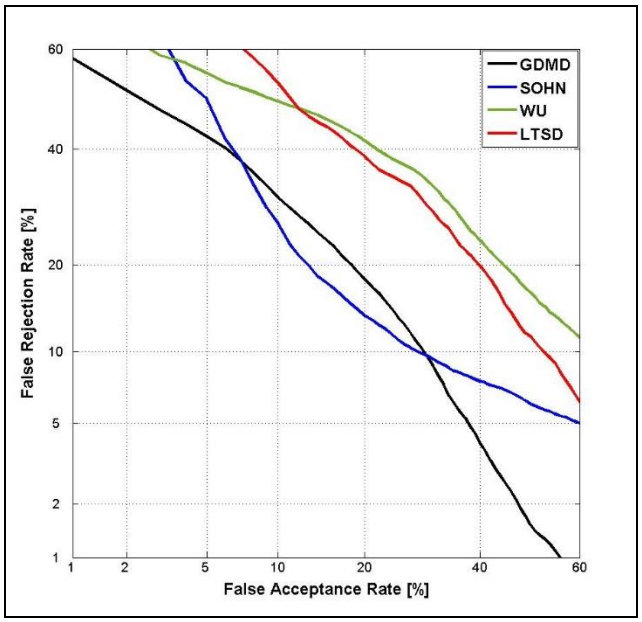
No.	Features	Errors	fixed1thr					
			0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
1	**GDMD**	**FEC**	1.9404	3.1314	4.6510	5.9814	7.3747	8.8881
		BEC	3.9413	5.6951	7.4600	9.1116	10.8529	12.2748
		MSC	2.9571	4.1935	5.3975	6.5006	7.3492	8.0552
		SDN	1.1903	2.3488	3.7620	5.5413	7.7321	10.5952
		FEA	8.8823	7.7410	6.7685	5.6769	4.8258	3.8666
		OVER	4.6757	4.2099	3.6801	3.3302	2.8368	2.0884
		NDS	2.6701	1.6557	1.2362	**0.9857**	0.9923	1.0133
2	**Wu**	**FEC**	0.8093	1.5296	2.3351	3.5578	4.8266	6.3039
		BEC	2.5580	4.0055	5.5531	7.1991	8.9174	10.5640
		MSC	1.6466	2.6775	4.0503	5.6328	7.3398	9.0469
		SDN	0.1780	0.4021	0.7464	1.1710	1.9155	3.1550
		FEA	8.2984	6.6272	5.4429	3.9902	3.1149	2.2396
		OVER	4.8943	3.5620	2.7683	2.2087	1.7109	1.1225
		NDS	5.6702	3.4814	2.2473	1.6645	1.4978	1.4680
3	**GDMD adapt1thr**	**FEC**	2.288503					
		BEC	4.273201					
		MSC	3.212371					
		SDN	1.432494					
		FEA	8.641757					
		OVER	4.471549					
		NDS	2.487996					
4	**LTSD**	**FEC**	0.0000					
		BEC	0.0024					
		MSC	0.0000					
		SDN	0.0000					
		FEA	7.3657					
		OVER	4.6017					
		NDS	2.4206					
5	**Sohn**	**fixed1thr**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**
		FEC	1.0670	1.4032	1.886957	2.3600	3.3343	4.2296
		BEC	0.6816	0.9562	1.359655	1.8116	2.3071	3.1438
		MSC	1.1305	1.3932	1.710775	2.0755	2.4796	2.9054
		SDN	0.1400	0.1687	0.235947	0.2957	0.3529	0.4488
		FEA	5.4881	4.7221	3.858933	3.1392	2.2782	1.7903
		OVER	9.1969	7.9220	6.844749	5.5808	4.5775	3.6216
		NDS	12.1772	11.4222	10.6087	10.0513	9.1572	8.2366

Таблица 4.11. Корпус NOIZEUS - Стойности на AUC за различни видове шум при SNR=5 dB

	Features	Airport	Babble	Car	Exhibition	Restaurant	Station	Street	Train
1	**log-GDMD**	0.8011	0.8103	0.8280	**0.8492**	0.8198	**0.8199**	**0.7890**	**0.8156**
2	**Sohn**	0.7806	0.776	0.7603	0.8295	0.8036	0.7703	0.7782	0.7763
3	**Wu**	0.7511	0.7885	0.8239	0.8341	0.7996	0.7973	0.7600	0.8016
4	**LTSD**	**0.8112**	**0.8119**	**0.8361**	0.8420	**0.8228**	0.8139	0.7633	0.7944

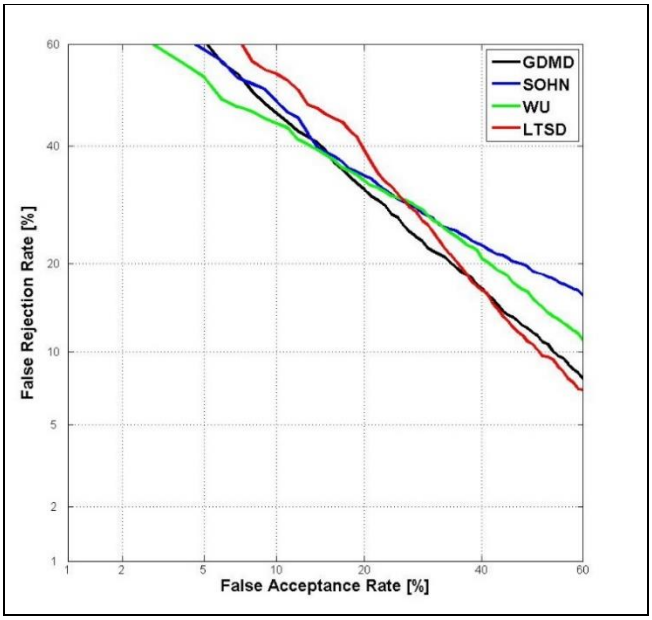
Таблица 4.12. Корпус DanEllis - Стойности на AUC

	Features	AUC
1	**log-GDMD**	**0.9068**
2	**Sohn**	0.8958
3	**Wu**	0.7802
4	**LTSD**	0.7988



Фиг. 4.5. DET графики за данни от корпуса на DanEllis и различни детектори.

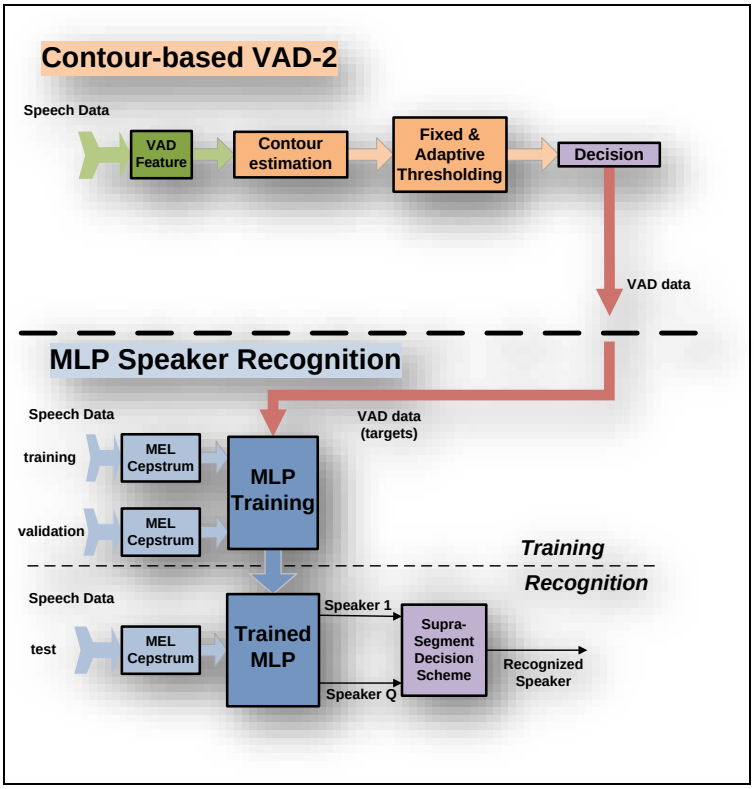
На фиг. 4.6 са показани средните DET-графики визуализиращи изменението на грешката при детекция на говор за данни от корпуса NOIZEUS при шум от движещ се влак със SNR=5 dB и изследваните тук детектори.



Фиг. 4.6. DET графики за данни от корпуса на NOIZEUS, шум от влак, SNR=5 dB и различни детектори.

4.9. Система за идентификация на диктори с VAD-2

Използваната при VAD-2 система за независима от текста идентификация на диктори е аналогична на тази разгледана в т. 4.6.2. Блоковата схема на системата за разпознаване включваща детайли за VAD-модула е показана на фиг.4.7.



Фиг. 4.7. Блокова схема на системата за разпознаване на диктори включваща детайли за модула VAD-2.

4.9.1. Експериментални резултати.

Стойностите на CEN, получените грешки (в проценти) при идентификация на диктори за различни признаци и прагови стойности с данни от корпуса BG-SRDat са показани в таблица 4.13.

Таблица 4.13. Корпус BG-SRDat - грешки при идентификация на диктори в проценти и стойности на CEN

No.	Features		fixed1thr				
			0.1	**0.2**	**0.3**	**0.4**	**0.5**
1	**log-GDMD**	**Err**	15.19	14.07	**13.88**	14.00	16.93
		CEN	0.1594	0.1437	**0.1436**	0.1463	0.1719
2	**Wu**	**Err**	19.87	**16.38**	19.16	18.04	20.74
		CEN	0.1933	**0.1718**	0.1832	0.1857	0.2026
3	**LTSD + HangETSI**	**Err**	17.79				
		CEN	0.2438				
4	**log-GDMD + adapt1thr**	**Err**	13.35				
		CEN	0.1432				
5	**Sohn**		**fixed1thr**				
			0.3	**0.4**	**0.5**	**0.6**	**0.7**
		Err	18.81	19.94	19.07	**17.63**	17.69
		CEN	0.1860	0.2002	0.1913	**0.1831**	0.1868

Резултатите в таблица 4.13 показват грешката от идентификация получена при различни прагови стойности. Ако се сравнят резултатите от детекцията на говор в таблица 4.9 и тези в таблица 4.13 се вижда, че грешката при разпознаване на диктори зависи силно от точността, с която се разпознават не-говорните сегменти. Минималната грешка при признака log-GDMD е при коефициент със стойност 0.3 (този коефициент формира фиксирания праг във ф-ла 3.20). В таблица 4.10 при тази стойност на коефициента и при същия параметър, за грешка NDS се наблюдава локален минимум. Тази грешка при детекция е шум класифициран като говор, т.е. добавени са фрагменти с шум при обучение на модела на диктора.

Интересен е резултатът, получен при използване на адаптивния праг, предложен от автора в гл. 3 – т. 3.4.2. При него се получава най-малка стойност на грешката (13.35%). Идеята на автора при формирането на този праг е той да се използва при разпознаване на диктори с кратки фрази, където често е налице низходяща интонация в края на произнасянето (респективно по-ниски нива в анализирания контур). В случая обаче прочетения текст е значително по-дълъг (> 30 секунди). При сравнение на грешките от детекция в таблица 4.10 между log-GDMD с фиксиран праг с coeff=0.3 и log-GDMD с адаптивен праг се вижда че при фиксирания праг грешките FEC, BEC, MSC и SDN (грешки от изрязване) са повече от два пъти по-големи от тези при адаптивния праг (clipping 27.13% vs 11.20%). И обратно грешките от вмъкване са около два пъти по-малки (insertion 12.73% vs 23.19%). Видно е в случая значителното изрязване на говорни фрагменти, което води до увеличаване на грешката при идентификация. Това увеличение на грешката в случая не се компенсира от по-точното разпознаване на не-говорните сегменти.

Изследванията описани в Глава 4 могат да се разглеждат като последващо развитие на идеите на автора отразени в [Ouzounov, 2006], [Ouzounov, 2009] и отнасящи се до разработване на MLP-VAD-алгоритми в системи за идентификация на диктори.

При експерименталните изследвания реализирани в дисертацията е използван Matlab ver. 2009 закупен по договор „УЕБ-базирана интерактивна система, подпомагаща построяването на модели и решаването на задачи за оптимизация и вземане на решения“, Договор с Фонд „Научни изследвания, ДТК 02 /71 от 17.12.2009 г.

4.10. Заключение

В настоящата глава е извършен сравнителен експериментален анализ на ефективността на предложените в гл. 2 признаци за детекция на говор. За всеки признак (референтен или предложен от автора) се формира отделен детектор на говор, който става част от система за независима от текста идентификация на диктори реализирана чрез невронна мрежа (многослоен перцептрон).

Реализирани са два детектора на говор – VAD-1 и VAD-2. Разработването на два отделни VAD модула се налага поради факта, че в гл. 2 са предложени два вида параметри. Едните са скаларни величини и са предназначени за детекция на говор чрез контурен анализ, докато другите са вектори и се използват като параметрични представяния в класификационни схеми.

При VAD-1 за детекция на говор се използва невронен класификатор с многослоен перцептрон. Локализираните от този модул говорни сегменти се подават на алгоритъма за разпознаване на диктори.

При VAD-2 се използва времеви контур и прагове (подобно на алгоритмите разгледани в гл. 3). За всеки сегмент стойността на контура се сравнява с фиксиран праг и се приема решение не-говор/говор. Впоследствие, аналогично на VAD-1, само говорните сегменти се подават на алгоритъма за разпознаване.

Оценката на ефективността на детекторите е реализирана в два етапа. На първия етап е оценена точността на детекция получена при различните детектори. Тя е определена чрез анализ на разликите между ръчно и автоматично (от съответния алгоритъм) локализираните сегменти не-говор/говор. Изчислени са няколко вида грешки всяка от които описва различни характеристики на VAD алгоритъма. Допълнително са изчислени и параметри, които служат за оценка на точността на бинарната класификация. На този етап са реализирани експерименти със зашумени говорни данни на български и английски език.

На втория етап е оценено влиянието на алгоритмите за детекция на говор върху точността на разпознаване в система за независима от текста идентификация на диктори. Тук експериментите са реализирани с говорни данни на български език записани по телефонен канал.

На базата на получените експериментални резултати са направени следните изводи:

- **Детектор VAD-1**

Грешката при идентификация на диктори както и ентропията CEN при детектора VAD-1 са с минимални стойности за предложените признаци BMD и MMD. Обаче при тестовете за точност получените резултати нямат еднозначно тълкуване – максималната стойност на F-measure е получена за параметъра MMD, докато максималната стойност на AUC е при параметъра IW-MFCC. Тези различия са свързани с особеностите на разпределението на данните в изхода на многослойния перцептрон използван в детектора и изискват по-обстоен анализ.

- **Детектор VAD-2**

Детектора на базата на предложения признак log-GDMD, използван във VAD-2 в повечето тестове превъзхожда тези на базата на алгоритмите на Sohn, Wu и LTSD. Необходимо е да се уточни, че признака LTSD е адаптивен спрямо променящото се ниво на шума, а в алгоритъма на Sohn е включена процедура за оценка на шума. При признака log-GDMD се разчита само на присъщата (вътрешна) робастност на неговите два компонента – на модифицирания спектър на групово закъснение и на делта спектралната автокорелационна функция. Тази робастност е обусловена от факта, че те се основават в голяма степен на свойствата на производните - от една страна на първата производна спрямо честотата на фазовия спектър и от друга на първата производна на спектралната автокорелационна функция.

- **VAD-1 vs VAD-2**

Сравнението между двата детектора от гледна тока на точността е трудно да бъде направено директно чрез стойностите на SHR и NHR. По-ефективен подход е да се сравнят стойностите на AUC. В този случай при VAD-1 параметрите BMD и MMD имат стойности на AUC съответно 0.9028 и 0.9043, докато при VAD-2 стойността на AUC за log-GDMD е 0.9143. Ако се анализира грешката, получена при идентификация то при VAD-2 за log-GDMD тя е 13.35/13.88% (в зависимост от прага), а при VAD-1 – 14.46/15.94% - съответно за BMD и MMD. Въз основа на посочените резултати се счита, че VAD-2 с log-GDMD е по-успешен детектор в контекста на реализираните тестове. Автора смята, че по-лошите резултати, получени при VAD-1 не са следствие от използваните параметри, които са във векторна форма и представят кратковременния спектър достатъчно детайлно. Основния проблем при този детектор е използването на две невронни мрежи в една обща система, което в повечето случаи води до трудности при тяхната настройка.

4.11. Резюме на получените резултати към Глава 4

Научно-приложни резултати:

1. Предложени са два алгоритъма за детекция на говорни сегменти условно означени като VAD-1 и VAD-2. При VAD-1 се използва невронен класификатор с многослоен перцептрон и признаци във векторна форма. При VAD-2 се използват признаци в скаларна форма и прагова логика (Глава 4, т.4.6 и т.4.8).
2. Направен е сравнителен експериментален анализ на ефективността на предложените в гл. 2 признаци спрямо избрани референтни такива при използването им във VAD-1 и VAD-2. Сравнението е на базата на грешките от детекция на сегментно ниво и на критерия за точност при бинарна класификация. Реализирани са експерименти със зашумени говорни данни на български и английски език (Глава 4, т.4.6.5 и т.4.8.4).
3. Направен е сравнителен експериментален анализ на ефективността на предложените в гл. 2 признаци спрямо избрани референтни такива при използването им във VAD-1 и VAD-2. Сравнението е на базата на грешката при разпознаване в система за независима от текста идентификация на диктори реализирана чрез класификатор с многослоен перцептрон. При експериментите са използвани говорни данни на български език записани по телефонен канал (Глава 4, т.4.7 и т.4.9).

ГЛАВА 5

BG-SRDat – Корпус с говорни данни записани по телефонен канал и предназначен за разпознаване на диктори

5.1. Увод

Известно е, че стандартните (сертифицирани) корпуси с организирани говорни данни са от изключителна важност при реализиране на изследвания в областта на разпознаване на диктори. Избора на корпуса и методиката на неговото използване определя насоката на изследванията, позволява да се демонстрират статистически значими резултати, както и да се сравняват резултати от различни работи. Създаването на даден корпус винаги е обусловено от задачите, които той трябва да решава. Например корпус, предназначен за разпознаване на диктори в криминалистиката ще се различава значително по структура и съдържание от корпус ориентиран към детекция на диктори (speaker labeling) в събрания, срещи (meetings) [Sturim et al., 2016].

В работата е описан корпуса BG-SRDat (Bulgarian language Speaker Recognition DATa) съдържащ говор записан по телефонен канал (стационарни и мобилни телефони и чрез VoIP) и включващ фрази и разговори на български и само фрази на английски език. Акцента при изграждане на корпуса е многообразието на комуникационната среда, т.е. различни телефонни канали, различно местоположение на диктора, различен съпътстващ шум при произнасяне на фразите и др. Корпуса съдържа 630 записа с различна продължителност, събрани от 40 диктора-мъже. Началната версия на този корпус е описана в [Ouzounov, 2003].

5.2. Общи параметри на корпусите с говорни данни

В литературата са описани изискванията, които трябва да бъдат изпълнени, за да се създаде т.н. добър корпус с говорни данни, предназначен за разпознаване на диктори [Sturim et al., 2016]. Едно от най-важните изисквания е данните да са достатъчно представителни така че заключенията направени при разпознаване на диктори с тези данни да са с висока степен на генерализация. Факторите, които влияят върху представителността на данните са:

- брой диктори – достатъчни за да се получат статистически значими резултати;

- брой сесии за диктор – най-малко 2 сесии – броят трябва да е достатъчен за да обхване изменчивостта в данните причинена от диктора, комуникационния канал и условията на реализация на говорното съобщение.
- дължина на записа – поне 30 секунди реален говор на сесия. Най-често дължината на записите е значително по-голяма като впоследствие тя се редуцира съобразно конкретните цели на изследването;
- времеви интервал между сесиите – достатъчен, за да се отчетат времевите вариации на анализираните говорни параметри;
- общ период от време при създаване на корпуса – достатъчно дълъг, за да се отчетат възрастовите изменения в говорните параметри;
- разнообразие на дикторите – достатъчно на брой и различни диктори, за да е възможно да се анализира влиянието върху процеса на разпознаване на фактори от вида – възраст, пол, месторождение, матерен език, социално икономически статус и др.

Освен изброените по-горе фактори трябва да се разгледат и факторите, които влияят върху двата вида изменчивост – присъща или вътрешна (intrinsic variability) и външна (extrinsic variability).

Факторите, които влияят на вътрешна изменчивост са:

- пол – обикновено много по-ефективно е разпознаването при използване на данни от един и същи пол;
- говорен стил – разговорен, по зададен текст (prompted) и интервю;
- сила на говора – номинално, шепот и публична реч (orated);
- език, диалект и разговорна реч;
- стрес, емоции, психично състояние;
- здравен статус - заболявания на гласовия тракт.

Факторите, които влияят на външната изменчивост са:

- комуникационен канал – микрофон, телефон и кодиране (4G, VoLTE, VoIP);
- среда на реализация – офис, движещо се превозно средство, фонов шум, съпътстващи разговори и др.

Всичките изброени по-горе фактори не действат самостоятелно, а са взаимно свързани и обусловени един от друг. Засега не е възможно да се създаде корпус, при който всеки един от факторите да бъде анализиран самостоятелно.

Необходимо е да се отбележи, че изпълнението на подобни изисквания (т.е. създаване на добър корпус) е възможно само при наличие на сериозни финансови и времеви ресурси.

5.3. Кратко описание на известни корпуси

В текста накратко ще бъдат описани някои от най-известните сертифицирани корпуси с говорни данни на английски език предназначени за разпознаване на диктори по телефонен канал. Повече информация за стандартизираните корпуси с говорни данни могат да бъдат в намерени в сайтовете на Linguistic Data Consortium (LDC) [LDC, online] и European Language Resources Association [ELRA, online]

5.3.1. SWITCHBOARD

Корпуса SWITCHBOARD-1 Release 2 съдържа 2430 телефонни разговора със средна продължителност на всеки от тях 6 минути, проведени от 543 диктори (302 мъже и 241 жени) - общо около 240 часа говор. За всеки разговор в базата от данни е включена и ортографичната му транскрипция. Обозначени са също така и неговорните акустични събития, смес от говор на двама диктори, прекъсвания и др. [SWITCHBOARD, online],

5.3.2. SPIDRE

Корпуса SPIDRE [SPIDRE, online] е получен чрез селекция на част от говорния материал, включен в SWITCHBOARD. Той съдържа 280 разговора, в 180 от които участва поне един от т.н. основни диктори (target speakers). Останалите 100 разговора са водени от неосновни диктори. В SPIDRE са включени 45 основни и 287 неосновни диктори. Основната цел при създаването на тази база от данни е да се даде възможност на изследователя да анализира влиянието на параметрите на телефонната апаратура върху точността на разпознаване.

5.3.3. TIMIT и негови варианти

Базата TIMIT [TIMIT, online] съдържа говорен материал, който включва три вида изречения - 2 диалектни, 450 фонетично балансирани и 1890 избрани от печатен текст (при максимално разнообразие на алофоните). Записана са 630 диктори всеки от които е произнесъл 10 изречения. Записите са реализирани в студио с висококачествен микрофон. Базата включва също и ортографична и фонетична транскрипция на говорния материал, както и времева сегментация в милисекунди.

На базата на TIMIT са създадени допълнителни корпуси (варианти на TIMIT) чрез възпроизвеждане на оригиналните записи и предаването на така получения говор по различни комуникационни канали. Някой от най-популярните варианти на TIMIT са:

- NTIMIT – използвани са различни телефонни линии - общо 250 на брой като половината от тях са междуградски [NTIMIT, online].
- CTIMIT - записите се възпроизвеждат като се реализира акустична връзка между високоговорител и мобилен телефон. Всяка сесия е с отделно телефонно обаждане, направено при различни условия на запис – движещ се автомобил с променлива скорост и в различна среда - град, село и др. [CTIMIT, online].
- FFMTIMIT – този корпус съдържа непубликувани записи от създаването на TIMIT получени от втори микрофон, разположен на значително разстояние от диктора. В тях е налице значителен нискочестотен шум, предизвикан от климатичната инсталация в използваното помещение [FFMTIMIT, online].
- WTIMIT - тук е използвана широколентова връзка между мобилните телефони обменящи записи от TIMIT [WTIMIT, online].

5.3.4. RSR2015

Корпуса с говорни данни RSR2015 [Larcher et al., 2014] е предназначен за зависимо от текста разпознаване на диктори. Той съдържа аудио записи от 298 диктора (142 жени и 156 мъже), всеки от тях записан в 9 сесии или общо корпуса съдържа говор с продължителност от 151 часа. Записите за направени в офис чрез използване на 6 мобилни устройства (смартфони и таблети). Данните са организирани в три групи. Първа група съдържа 30 изречения от базата данни TIMIT с единична дължина 3.2 секунди и обща продължителност на записите - 71 часа. Втората група включва 30 кратки команди с единична дължина 2 секунди и обща продължителност - 45 часа. Третата група съдържа произнасяния на последователности от цифри съответно 3 последователности в 10 сесии и 10 последователности в 5 сесии.

5.3.5. NIST SRE 2018

Едни от най-известните корпуси в областта на разпознаване на диктори е създаден от Националния институт по стандарти и технология (National Institute of Standards and Technology - NIST). Значим факт в случая е, че освен данните се предлага цялостна методика за оценка на резултатите от разпознаване. Тя се формулира в така наречените оценъчни сесии – Speaker Recognition Evaluations (SRE) [NIST, online]. Последната сесия е SRE 2018 [Sadjadi et al., 2019]. Задачата при нея е speaker detection – определяне дали даден целеви диктор говори в анализирания тестов запис. Говорните данни в случая са свободни разговори по телефон, но записани в държави извън Северна Америка. Допълнителни при SRE 2018 са въведени два нови типа данни - записи от VoIP и аудио данни, извлечени от аматьорски онлайн видео записи (AfV).

За SRE18 са необходими три вида данни – за обучение, за настройка и за тестване. Условията при обучение могат да бъдат ограничени (fixed) и неограничени (open). При ограничени условия се използват предварително специфицирани данни от предходните сесии (SRE 1996-2016) и корпуси MIXER 6 [Mixer 6, online], SWITCHBOARD [SWITCHBOARD, online] и Fisher [Fisher, online] от LDC [LDC, online]. Допълнително могат да се използват VoxCeleb [VoxCeleb, online] и SITW [SITW, online]. При неограничени условия са избрани данни от IARPA Babel Program също достъпни чрез LDC [LDC, online].

Данните за настройка и тестване също са избрани от LDC [LDC, online] и са корпусите CMN2 и VAST [VAST, online]. Във втория корпус се съдържат AfV записи на английски език, получени от мобилни телефони. Тези разговори съдържат голямо разнообразие от шумови компоненти (съпътстващи разговори, смях, шум от трафик и др.) тъй като реализирани в реална среда (real-world environments).

Като цяло тестовите данни включват: при телефонни разговори – 188 диктора, тестове с целеви диктори – 19298 и с не-целеви – 2002232. При AfV данни – 101 диктора, тестове с целеви диктори – 315 и с не-целеви – 31500.

Основните заключения от SRE 2018 са две - първо, разпознаването на диктори чрез AfV данни е много по-трудно отколкото при използване на данни получени чрез стандартни телефонни разговори и второ значително подобрение на точността при разпознаване се получава чрез прилагане на сложни модели при представяне на данните както и при по-ефективно използване на данните за настройка.

5.4. Описание на BG-SRDat

При описание на корпуса ще бъде обърнато внимание на следните фактори: вида на говорния материал; брой сесии и периода между тях; условия, при които са реализирани записите и файлова структура.

5.4.1. Вид на говорния материал

Съобразно вида на говорния материал корпуса BG-SRDat може да се разглежда като съставен от 5 модула, които условно ще бъде означен като Speech Data (SD), съответно SD1, SD2, SD3, SD4 и SD5. Тези модули са:

- SD1 (BG) – съдържа записи на прочетен текст (с продължителност от около 40 секунди), избран от вестник. Реализирани са 2 вида записи на един и същ текст, съответно чрез микрофон (26 диктора с общо 28 записа) и по телефонен канал (30 диктора с общо 60 записа). 26 от дикторите са идентични в двата записа;

- SD2 (BG) – съдържа запис на кратка фраза. Реализирани са 373 записа от 20 диктора чрез стационарен и мобилен телефон. Фразата е: *“Здравей Манолов! Как се чувстваш днес?”.* Необходимо е да се доуточнят някои детайли за използваната фраза. Тя е предложена от автора поради факта, че съдържа преобладаващ брой съгласни. Фразата включва общо 31 фонеми, от които 10 гласни и 21 съгласни (от които 8 шумови беззвучни) и освен това започва с две шумови звучни съгласни *‘зд‘* и завършва с шумова беззвучна съгласна *‘с‘*. При запис от телефонен канал тези две особености на фразата затрудняват, както определянето на граничните точки, така и самото разпознаване на диктори.
- SD3 (BG) - съдържа запис на прочетен текст (със средна продължителност от около 80 секунди) от вестник. С цел да се постигне в някаква степен лексикално разнообразие са избрани различни по тематика дописки. Реализирани са 14 записи от 10 диктора чрез стационарен и мобилен телефон;
- SD4 (BG) – включени са разговори на свободна тема, като в някои фрагменти едновременно разговарят и двамата диктори. Максимална продължителност на разговорите е около 7 минути. Реализирани са 4 записа от 4 диктора чрез мобилен телефон и VoIP.
- SD5 (EN) - съдържа запис на кратка фраза на английски език. Реализирани са 150 записа от 9 диктора чрез стационарен телефон.

5.4.2. Брой сесии и период между тях

- SD1 - реализирани са поне 2 сесии за диктор като всяка сесия съдържа само един запис. Периода между сесиите е около 3 месеца.
- SD2 - тук говорния материал съдържа поне 10 сесии за диктор с не по-малко от 2 записа на сесия. Във всяка сесия записите са направени в един ден, но обажданията са от телефонни номера разположени на различни места в София и страната (при стационарните телефони). Периода между сесиите е около седмица.
- SD3 - за някой от дикторите са реализирани 2 сесии всяка от които включва по един запис с различен текстов материал. Периода между сесиите е около седмица.
- SD4 - тук има само по един запис на сесия.
- SD5 - методиката е аналогична на тази използвана при SD2.

5.4.3. Условия, при които са реализирани записите

Целта, която е набелязана при реализацията на корпуса е да се постигне многообразие на условията, при които са реализирани записите. В резултат, на което дикторите правят

телефонни обаждания от различни места – тих/шумен офис, зали, телефонни автомати разположени на шумни улици и др. При записи чрез мобилни телефони са използвани различни модели апарати и дикторите в повечето случаи се движат по булеварди или магистрали с интензивен трафик.

5.4.4. Файлова структура

Името на файла за всеки запис е във вида: *'s_xxyyzz.wav',* където:

- *s* – тип говорни данни SD1,…,SD5 – съответно *'1',...., '5';*
- *xx* – идентификационния номер на диктора (ID) – започва от *'01'*;
- *yy* – номера на сесията, започва от *'01'*;
- *zz* – номера на произнасянето в текущата сесия, започва от *'01'*;

Например файл с име *'2_110501.wav'* означава, че записът е с данни SD2, за диктор с ID=11, пета сесия и първо произнасяне в сесията. Пълното име на даден файл в системата е *'#speaker/#speech_data/#session/utterance_name.wav'* или в посочения случай името на файла има вида *'11/SD2/05/2_110501.wav'.* За всеки диктор е създадена таблица описваща неговите файлове наречена Utterances Description Table (UDT). Ако той има записи в различните групи данни, то за всяка група се създава отделна негова таблица. Следните данни са включени в таблицата:

- ID на диктора;
- група говорни данни – SD1,….,SD5;
- място, от където се провежда разговора - офис/улица/булевард/магистрала/ движеща се кола и др.;
- вид телефонен разговор – локален/междуградски/международен/мобилен/VoIP;
- маркер за наличие на шум;

Решението за наличие на шум в записа и определяне на неговия вид е предмет на субективна оценка. Тук е въведен маркер (да/не) който има ориентировъчен характер. В таблицата се отбелязват следните два вида шум – от комуникационния канал (фонов, импулсен и смесване на разговори) и от средата, в която диктора говори (съпътстващи разговори, музика и шум от трафик). Тези видове шум в таблицата са означени по следния начин:

- BN – фонов шум (background noise);
- PN - импулсен шум (pulse noise);
- CT - смесване на разговори (cross-talks);
- BC - съпътстващи разговори (background conversations);
- M – музика (music);

- TN – шум от трафик (traffic noise);
- */- – маркер за наличие на шум (да/не);

В таблица 5.1 е показана като илюстрация UDT за диктор с ID=01 и SD2.

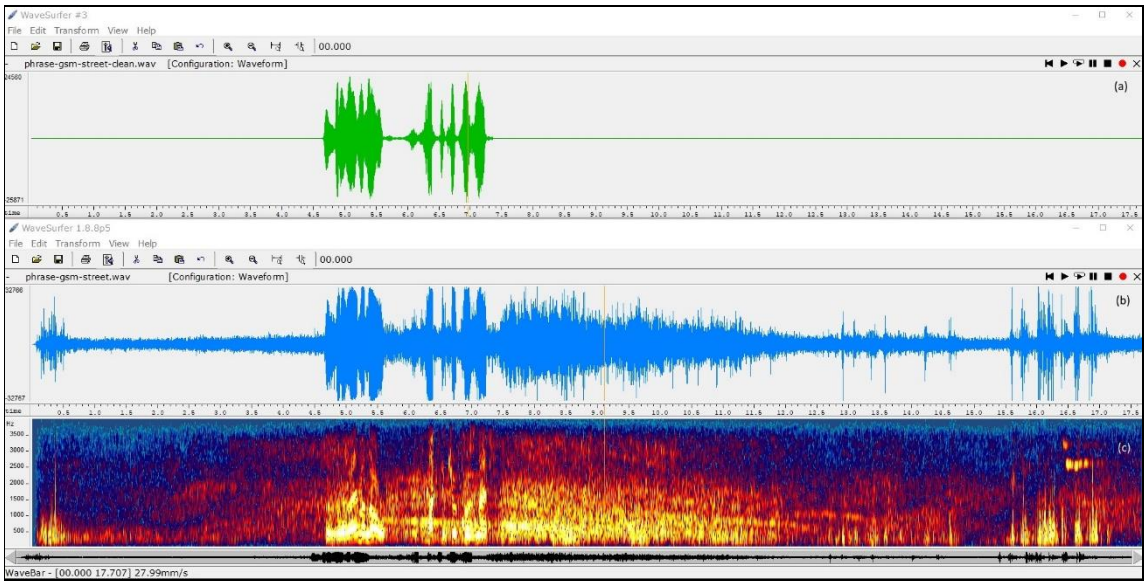
Таблица 5.1. UDT за диктор с ID=01 и SD2

Speech Data 2 ***Speaker ID = 01***										
№	***File name***	***Session/ Utterance***	***Calling place***	***Phone call***	***Noise presence marks***					
					Operation noise			***At speaker position***		
					BN	***PN***	***CT***	***BC***	***M***	***TN***
1	2_010101.wav	1/1	Street	Landline	*	-	-	-	-	-
2	2_010201.wav	2/1	Street	Landline	*	-	-	-	-	-
3	2_010301.wav	3/1	Office	Landline	*	-	-	-	-	-
4	2_010302.wav	3/2	Street	Landline	*	-	-	-	-	-
5	2_010401.wav	4/1	Street	Landline	*	*	-	-	-	-

Таблицата съдържа част от метаданните, които се въвеждат за всеки файл. Понастоящем поради сравнително немалкия обем на корпуса и перспективите за разрастването му със записи, получени от мобилни телефони, неговата структура постепенно се въвежда в MySQL база данни.

На фиг. 5.1 е показан необработен запис на фразата от SD2 получен чрез мобилен телефон и диктор намиращ се на улица с интензивен трафик. На фиг. 5.1 (a) е показана чистата версия на записа получена чрез специализирана програма, на фиг. 5.1 (b) е оригиналната зашумена версия, а на фиг. 5.1 (c) е спектрограма на зашумената версия.

На фиг. 5.1 (b) се вижда, че в интервала между 7.5 и 12 секунди е наличен шум от трафик (спиращ трамвай), докато в края на записа се наблюдава импулсен шум, причинен от рязко изместване на мобилния телефон в ръката на диктора. За получаване на графиката на фиг. 5.1 са използвани WaveSurfer [WaveSurfer, online] и Paint.Net [Paint.net, online].



Фиг. 5.1. Фразата от SD2 записана чрез мобилен телефон, когато диктора се намира на улица с интензивен трафик, съответно: a) чиста версия на записа; b) оригинална зашумена версия и c) спектрограма на зашумената версия.

5.5. Приложение на BG-SRDat

Корпуса е използван за изследвания в областта на верификация на диктори чрез фиксирани фрази (на български и английски), независима от текста идентификация на диктори и детекция на говор. Основна тенденция при бъдещо развитие на корпуса ще бъде постепенното му преобразуване в корпус съдържащ говорни данни получени само от мобилни устройства.

Приложни аспекти

Използваните в Глава 3 две приложения за верификация на диктори – DTW и HMM, както и първоначалната версия на алгоритъма за определяне на гранични точки са част от приложно-програмен интерфейс Speaker Recognition API (SR-API). Този интерфейс е разработен през 2003 г. от Fadata, Ltd., подизпълнител на фирма от САЩ, като автора е научен консултант на проекта. Подобрена версия на посочения алгоритъм е представена от автора за първи път в негова публикация от 2014 г.

Описания в Глава 5 корпус BG-SRDat е създаден изцяло на базата на идеите и методиката на автора. Първата версия на корпуса, която включва записи от стационарни телефони на български и английски език е разработена от AdVoice Ltd. Записите от мобилни телефони и VoIP са направени по-късно. Краткото описание на корпуса е включено в дисертацията само с информационна цел - читателите да придобият представа за характера на данните използвани в експериментите.

Резюме на получените резултати

Научни резултати:

1. Предложен е метод, при който чрез прилагане на делта-филтър върху спектралната автокорелационна функция е получена т.н. делта спектрална автокорелационна функция. Анализирани са особеностите на тази филтрация, при която е постигнато усилване в честотната област на хармоничната структура на говорния сигнал. (Глава 2, т. 2.1.3).
2. Предложен е подход за изчисляване на признаци за детекция на говор базиращ се на свойствата на делта спектралната автокорелационна функция. Чрез този подход са дефинирани три признака. Първия от тях (т.н. MD-признак) е в скаларна форма и е предназначен за детекция чрез анализ на времеви контури, докато другите два (т.н. BMD- и MMD - признаци) са вектори и са предназначени за детекция чрез алгоритми за разпознаване (Глава 2, т. 2.1.4).
3. Извършен е теоретичен анализ на изменението на спектъра на групово закъснение при зашумени с адитивен шум говорни сигнали. Този анализ е реализиран косвено, чрез изследване изменението на аргументите на проекционните функции на сходство на основата на адитивния спектрален модел (Глава 2, т. 2.2.3).
4. Предложен е подход за изчисляване на признаци за детекция на говор базиращ се комбинация на модифицирания спектър на групово закъснение и делта спектралната автокорелационна функция. Чрез този подход са дефинирани два признака (т.н. lin-GDMD и log-GDMD - признаци) които са предназначени за детекция на говор чрез анализ на времеви контури (Глава 2, т. 2.2.4).
5. Предложен е подход за определяне на гранични точки на говорно съобщение включващ алгоритъм за изчисляване на адаптивни прагови стойности и детерминиран краен автомат (Глава 3, т. 3.4.2-3).

Научно-приложни резултати:

1. Направен е сравнителен експериментален анализ на предложените в гл. 2 признаци спрямо избрани референтни такива. Сравнението е на базата на Евклидовото разстояние между Z-нормализирани времеви контури, изчислени за всеки признак съответно от чист и от зашумен сигнал (Глава 3, т. 3.3).

2. Разработени са три алгоритъма за определяне на гранични точки базиращи се на предложения подход и формирани съобразно използваните времеви контури (Глава 3, т. 3.5).
3. Направен е сравнителен експериментален анализ на ефективността на предложените в гл. 2 признаци спрямо избрани референтни такива при използването им в разработените алгоритми за определяне на гранични точки. Сравнението е на база точността на детекция оценена чрез хистограмен анализ на разликите между ръчно определените и получените от съответния алгоритъм гранични точки. Експериментите са реализирани със зашумени говорни данни на български и английски език (Глава 3, т. 3.6.3).
4. Направен е сравнителен експериментален анализ на ефективността на предложените в гл. 2 признаци спрямо избрани референтни такива при използването им в предложените алгоритми за определяне на гранични точки. Сравнението е на базата на грешката при разпознаване в две системи за зависима от текста верификация на диктори базирани съответно на DTW и HMM алгоритми. При експериментите са използвани говорни данни на български език записани по телефонен канал (Глава 3, т.3.6.4).
5. Разработени са два алгоритъма за детекция на говорни сегменти условно означени като VAD-1 и VAD-2. При VAD-1 се използва невронен класификатор с многослоен перцептрон и признаци във векторна форма. При VAD-2 се използват признаци в скаларна форма и прагова логика (Глава 4, т.4.6 и т.4.8).
6. Направен е сравнителен експериментален анализ на ефективността на предложените в гл. 2 признаци спрямо избрани референтни такива при използването им във VAD-1 и VAD-2. Сравнението е на базата на грешките от детекция на сегментно ниво и на критерия за точност при бинарна класификация. Реализирани са експерименти със зашумени говорни данни на български и английски език (Глава 4, т.4.6.5 и т.4.8.4).
7. Направен е сравнителен експериментален анализ на ефективността на предложените в гл. 2 параметри спрямо избрани референтни такива при използването им във VAD-1 и VAD-2. Сравнението е на базата на грешката при разпознаване в система за независима от текста идентификация на диктори реализирана чрез класификатор с многослоен перцептрон. При експериментите са използвани говорни данни на български език записани по телефонен канал (Глава 4, т.4.7 и т.4.9).

Заключение и идеи за бъдеща работа

В дисертационния труд е предложен метод за изчисляване на т.н. делта спектрална автокорелационна функция, която се получава чрез прилагане на делта-филтър върху спектралната автокорелационна функция. Предложени са два подхода за изчисляване на признаци за детекция на говор. При първия признаците се определят само чрез свойствата на делта спектралната автокорелационна функция. При втория подход те се получават чрез комбиниране на свойствата на делта спектралната автокорелационна функция и тези на модифицирания спектър на групово закъснение. Или общо са предложени пет признака - MD, BMD, MMD, log-GDMD и lin-GDMD. Те са използвани в предложени от автора алгоритми за определяне на гранични точки (ED-алгоритми) и алгоритми за детекция на говорни фрагменти (VAD-алгоритми). Тези алгоритми са включени в детектори на говор, които са част от системи за зависимо и независимо от текста разпознаване на диктори. Експериментално е сравнена ефективността на описаните детектори с тази получена при детектори използващи алгоритми с референтни признаци. Сравнението е извършено на два етапа. На първия етап е сравнена постигната точност на локализация на граничните точки и говорните фрагменти, а на втория е сравнена грешката, получена при разпознаване на диктори. При алгоритмите за определяне на гранични точки, използвани в системи за верификация на диктори чрез фиксирани фрази, доминиращ от гледна точка на точност на локализация и на минимална грешка при верификация е признакът log-GDMD. В системите за независима от текста идентификация на диктори минимална грешка при разпознаване е получена при използване на VAD-алгоритми съответно с BMD- и log-GDMD-признаци.

Бъдещата работа в областта на детекция на говор ще бъде насочена към разработка на хибридни VAD-алгоритми. При тях се комбинират различни признаци в един VAD-алгоритъм или се комбинират различни VAD-алгоритми в един общ класификатор.Този подход дава възможност за по-голяма адаптивност на детектора на говор при промяна в условията на средата.

Декларация за оригиналност на резултатите

Декларирам, че представената дисертация съдържа оригинални резултати, получени при проведени от мен научни изследвания. Резултатите, които са получени, описани и/или публикувани от други учени, са надлежно и подробно цитирани в библиографията. Настоящата дисертация не е прилагана за придобиване на научна степен в друго висше училище, университет или научен институт.

Подпис:

ПУБЛИКАЦИИ

По темата на дисертационния труд

1. **Ouzounov A.**, Cepstral Features and Text-Dependent Speaker Identification - A Comparative Study, *Cybernetics and Information Technologies,* vol. 10, No. 1, 2010, pp. 1-12. (***Реферирана в Web of Science***).
2. **Ouzounov A.,** Telephone Speech Endpoint Detection using Mean-Delta Feature, *Cybernetics and Information Technologies,* vol. 14, No. 2, 2014, pp. 127-139; (***Реферирана в SCOPUS***, ***SJR=0.138*****)**.
3. **Ouzounov A.,** Noisy Speech Endpoint Detection Using Robust Feature, Springer International Publishing Switzerland 2014, V. Cantoni et al. (Eds.): BIOMET 2014, LNCS 8897, pp. 105–117; (***Реферирана в SCOPUS, SJR=0.252*****)**.
4. **Ouzounov A.**, Mean-Delta Features for Telephone Speech Endpoint Detection, In: Proc. of the International Conference Automatics & Informatics, 2015, pp.185-188.
5. **Ouzounov A.,** Mean-Delta Features for Telephone Speech Endpoint Detection, Information Technologies and Control, No. 3-4, 2014, pp. 36-43.
6. **Ouzounov A.,** LTSD and GDMD features for Telephone Speech Endpoint Detection, *Cybernetics and Information Technologies,* vol. 17, No. 4, 2017, pp. 114-133; (***Реферирана в SCOPUS, SJR=0.204*****)**.

Цитирания на публикациите по темата на дисертацията

Цитирана статия:

Ouzounov A., Cepstral Features and Text-Dependent Speaker Identification - A Comparative Study, Cybernetics and Information Technologies, vol. 10, No. 1, 2010, pp. 1-12. (Реферирана в Web of Science)

Цитирания:

1. Mishra P., Agrawal S., Recognition Of Voice Using Mel Cepstral Coefficient & Vector Quantization, *International Journal of Engineering Research and Applications (IJERA,)* Vol. 2, Issue 2, Mar-Apr 2012, pp.933-938.: ISSN: 2248-9622.
   http://www.ijera.com/papers/Vol2_issue2/FA22933938.pdf

2. Mishra P., Agrawal S., Recognition of Speaker Using Mel Frequency Cepstral Coefficient & Vector Quantization, *International Journal of Science, Engineering and Technology Research (IJSETR*), Volume 1, Issue 6, December 2012, pp.12-17, ISSN: 2278 – 7798.
   http://ijsetr.org/wp-content/uploads/2013/08/IJSETR-VOL-1-ISSUE-6-12-17.pdf

3. Mishra P., Agrawal S., Recognition of Speaker Using Mel Frequency Cepstral Coefficient & Vector Quantization for Authentication, *International Journal of Scientific & Engineering Research (IJSER),* Volume 3, Issue 8, August 2012, pp. 1-6, ISSN 2229-5518.
   https://www.ijser.org/onlineResearchPaperViewer.aspx?Recognotion-of-Speaker-Useing-Mel-Cepstral-Coefficient-Vector-Quantization-for-Authentication.pdf

4. Бакина И. Г., Морфологическое сравнение изображений гибких объектов на основе циркулярных моделей при биометрической идентификации личности по форме ладони, Диссертация - кандидат физико-математических наук, Московский государственный

университет имени М. В. Ломоносова, 2011. Научная библиотека диссертаций и авторефератов disserCat:
http://www.dissercat.com/content/morfologicheskoe-sravnenie-izobrazhenii-gibkikh-obektov-na-osnove-tsirkulyarnykh-modelei-pri

5. Jain A. and O. Sharma, A Vector Quantization Approach for Voice Recognition Using Mel Frequency Cepstral Coefficient (MFCC): A Review, *International Journal of Electronics & Communication Technology*, vol. 4, Issue SPL-4, 2013, pp. 6-29, ISSN: 2230-7109 (Online), | ISSN: 2230-9543 (Print).
http://www.iject.org/vol4/spl4/c0128.pdf

6. Chen Y., E. Heimark, D. Gligoroski, Personal Threshold in a Small Scale Text-Dependent Speaker Recognition, International Symposium on Biometrics and Security Technologies (ISBAST), July, 2013, pp.162-170, DOI: 10.1109/ISBAST.2013.29, Print ISBN: 978-0-7695-5010-7. Publisher:IEEE.
http://ieeexplore.ieee.org/xpl/abstractReferences.jsp?tp=&arnumber=6597684&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6597684

7. Kumar, R.C.P., D.A. Chandy, Audio retrieval using timbral feature, International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN), March 25-26, 2013, pp. 222-226, DOI: 10.1109/ICE-CCN.2013.6528497, Print ISBN: 978-1-4673-5037-2. Publisher: IEEE.
http://ieeexplore.ieee.org/xpl/abstractReferences.jsp?tp=&arnumber=6528497&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6528497

8. Thakur A., R. Kumar, A. Bath, and J. Sharma, Automatic Control of Instruments Using Efficient Speech Recognition Algorithm, International Journal of Electrical & Electronics Engineering, Vol. 1, Spl. Issue 1, March, 2014, pp.16-22, e-ISSN: 1694-2310, p-ISSN: 1694-2426.
http://ijeee-apm.com/Uploads/Media/Journal/20140328142123_ACCT_14_IG_52.pdf

9. Kumar C. P. R., S. Suguna and J. Becky Elfreda, Audio Retrieval based on Cepstral Feature, *International Journal of Computer Applications* 107(8):28-33, December 2014. DOI: 10.5120/18774-0079; ISSN 0975 – 8887.
https://research.ijcaonline.org/volume107/number8/pxc3900079.pdf

10. Kamil I., K. Oyeyiola, Comparative Study on the Performance of Mel-Frequency Cepstral Coefficients and Linear Prediction Cepstral Coefficients under different Speaker's Conditions, *International Journal of Computer Applications*, March 2014, vol. 90, No. 11, pp. 38-42. DOI: 10.5120/15767-4460; ISSN 0975 – 8887.
https://research.ijcaonline.org/volume90/number11/pxc3894460.pdf

11. R. Christopher Praveen Kumar and S. Suguna, Analysis of MEL based features for audio retrieval, *ARPN Journal of Engineering and Applied Sciences*, Vol. 10, No. 5, March 2015, pp.2167-2171. ISSN 1819-6608.
http://www.arpnjournals.com/jeas/research_papers/rp_2015/jeas_0315_1735.pdf

12. Patil C. and G. Dhoot, A Security System by using Face and Speech Detection, International Journal of Current Engineering and Technology, Vol. 4, No. 3 (June 2014), pp. 2176-2182; E-ISSN 2277 – 4106, P-ISSN 2347 – 5161.
https://inpressco.com/wp-content/uploads/2014/06/Paper1922176-2182.pdf

13. Jain, A., Sharma, O.P., Evaluation of MFCC for speaker verification on various windows, Int. Conf. on Recent Advances and Innovations in Engineering (ICRAIE), 2014, pp.1-6, Print ISBN:978-1-4799-4041-7; DOI:10.1109/ICRAIE.2014.6909144; Publisher: IEEE https://ieeexplore.ieee.org/document/6909144

14. Abdelmajid, L., K. Mohamed, Extracting Multi Band Approach of Acoustic Vectors Extractors: Using HMM Classifier, International Journal of Science Technology & Engineering, Vol. 3, No.2, 2016, pp. 305-310; ISSN (online): 2349-784X. https://www.ijste.org/articles/IJSTEV3I2095.pdf

15. Bakina, I., Person recognition by hand shape based on skeleton of hand image, Pattern Recognition and Image Analysis, 2011, vol. 21, issue 4, pp. 694-704, Springer, Print ISSN 1054-6618,Online ISSN 1555-6212, DOI: 10.1134/S1054661811040031. https://link.springer.com/article/10.1134/S1054661811040031

16. Trabelsi I., M. Bouhlel, Learning vector quantization for adapted Gaussian mixture models in automatic speaker identification, Journal of Engineering Science and Technology Vol. 12, No. 5 (2017) 1153 – 1164. ISSN: 1823-4690. http://jestec.taylors.edu.my/Vol%2012%20issue%205%20May%202017/12_5_1.pdf

17. Sharma R., R. Bhukya, S. Prasanna, Analysis of the Hilbert Spectrum for Text-Dependent Speaker Verification, Speech Communication, Elsevier B.V., vol. 96, 2018, pp. 207-224. https://doi.org/10.1016/j.specom.2017.12.001, ISSN 0167-6393.

18. Кралева, Р., Разпознаване на реч: Корпус от говорима детска реч на български език, Университетско издателство „Неофит Рилски“, 2019; ISBN: 978-954-00-0199-9. https://www.researchgate.net/publication/335739119_Razpoznavane_na_rec_Korpus_ot_govorima_detska_rec_na_blgarski_ezik

Цитирана статия:

Ouzounov A., Telephone Speech Endpoint Detection using Mean-Delta Feature, Cybernetics and Information Technologies, vol. 14, No. 2, 2014, pp. 127-139, (***Реферирана в SCOPUS, SJR=0.138*****).**

Цитирания:

1. Guo Yu, Zhang Erhua, Liu Chi, An endpoint detection algorithm based on frequency-domain characteristics and transition fragment judgment, Journal of Shandong University (Engineering Science), 2016, Vol. 46 Issue (2), pp. 57-63; DOI: 10.6040/j.issn.1672-3961.2.2015.147. https://caod.oriprobe.com/articles/48238509/An_endpoint_detection_algorithm_based_on_frequency.htm

2. Sopon P., J. Polpinij, T. Suksamer, Speech-Based Thai Text Retrieval, The 11th National Conference on Computing and Information Technology, NCCIT'2015, pp. 259-264. http://202.44.34.144/nccitedoc/admin/nccit_files/NCCIT-20150810110354.pdf

3. Li, L., Y.Wang, X.Li, An Improved Wavelet Energy Entropy Algorithm for Speech Endpoint Detection, Journal of Computer Engineering, 2017, vol. 43, No. 5, pp. 268-274, DOI:10.3969/j.issn.1000-3428.2017.05.043; ISSN: 1000-3428.
http://manu55.magtech.com.cn/Jwk_ecice/EN/abstract/abstract27746.shtml#

4. Roy T., T.Marwala, S. Chakraverty, Precise detection of speech endpoints dynamically: A wavelet convolution based approach, Communications in Nonlinear Science and Numerical Simulation, Elsevier B. V., 2019, vol. 67, pp. 162-175; ISSN: 1007-5704.
https://doi.org/10.1016/j.cnsns.2018.07.008

5. Zhang, X., Q. Xiong, Y. Dai and X. Xu, Voice Biometric Identity Authentication System Based on Android Smart Phone, 2018 IEEE 4th International Conference on Computer and Communications (ICCC), Chengdu, China, Dec. 2018, pp. 1440-1444; Electronic ISBN:978-1-5386-8339-2; USB ISBN:978-1-5386-8338-5; Print on Demand(PoD) ISBN:978-1-5386-8340-8.
https://doi.org/10.1109/CompComm.2018.8780990;

6. Li, Y., C. L.-F. Cheng, P.-L. Zhang, Speech Endpoint Detection based on Improved Spectral Entropy, Journal of Computer Science, 2016, vol.43, No.11A, pp.233-236; ISSN 1002-137X.
http://journal.jsjkx.com/jsjkx/ch/reader/view_abstract.aspx?file_no=201611A053&flag=1

Цитирана статия:

Ouzounov A., Noisy Speech Endpoint Detection Using Robust Feature, Springer International Publishing Switzerland 2014, V. Cantoni et al. (Eds.): BIOMET 2014, LNCS 8897, pp. 105–117, (Реферирана в SCOPUS, SJR=0.252).

Цитирания:

1. Li, Y., C. L.-F. Cheng, P.-L. Zhang, Speech Endpoint Detection based on Improved Spectral Entropy, Journal of Computer Science, 2016, vol. 43, No. 11A, pp. 233-236; ISSN 1002-137X.
http://journal.jsjkx.com/jsjkx/ch/reader/view_abstract.aspx?file_no=201611A053&flag=1

Общо 25 цитирания – 24 в чужбина и 1 в България.

БИБЛИОГРАФИЯ

1. Abdulla·, W., Z. Guan, H. Sou, Noise Robust Speech Activity Detection, IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2009, pp. 473-477.
2. Akant, K., R. Pande, S. Limaye, Accurate Monophonic Pitch Tracking Algorithm for QBH and Microtone Research, *The Pacific Journal of Science and Technology*, 2010, vol. 11. No. 2, pp. 342-352.
3. Alam, M., P. Kenny, P. Ouellet, T. Stafylakis, P. Dumouchel, Supervised/Unsupervised Voice Activity Detectors for Text dependent Speaker Recognition on the RSR2015 Corpus, Odyssey 2014: The Speaker and Language Recognition Workshop, pp. 123-130.
4. Ashihara, T., Y. Shinohara, H. Sato, T. Moriya, K. Matsui, T. Fukutomi, Y. Yamaguchi, Y. Aono, Neural Whispered Speech Detection with Imbalanced Learning. Proc. Interspeech, 2019, pp. 3352-3356.
5. Beigi, H., Fundamentals of Speaker Recognition, Springer Science, 2011.
6. Bengio, S., J. Mariethoz, A Statistical Significance Test for Person Authentication, In: Proc. of ODYSSEY, The Speaker and Language Recognition Workshop, 2004, pp. 237-244.
7. Buyuk, O., M. Arslan, Model Selection and Score Normalization for Text-Dependent Single Utterance Speaker Verification, Turkish Journal of Electrical Engineering & Computer Sciences, vol. 20, 2012, No. Sup. 2, pp. 1277-1295.
8. Campbell, W., D. Sturim, D. Reynolds, Support Vector Machines Using GMM Supervectors for Speaker Verification IEEE Signal Processing Letters, 2006a, vol. 13, No. 5, pp. 308-311.
9. Campbell, W., D. Sturim, D. Reynolds, A. Solomonoff, SVM based speaker verification using a GMM supervector kernel and NAP variability compensation, ICASSP 2006b, pp. 97-100.
10. Cao, D., X, Gao, L. Gao, An Improved Endpoint Detection Algorithm Based on MFCC Cosine Value. Wireless Personal Communications, 2017, vol. 95, pp. 2073–2090.
11. Chen, L. M. Ozsu, V. Oria, Robust and Fast Similarity Search for Moving Object Trajectories, SIGMOD '05: Proceedings of the ACM SIGMOD international conference on Management of data, 2005, pp. 491–502.
12. Chung, H., S. J. Lee, Y. K. Lee, Weighted Finite State Transducer–Based Endpoint Detection Using Probabilistic Decision Logic, ETRI Journal, 2014, 36, pp. 714-720.
13. Fisher English Training Speech Part 1, Linguistic Data Consortium (LDC), https://catalog.ldc.upenn.edu/LDC2004S13, last accessed August 2019.
14. Comas, C., E. Monte-Moreno, J. Solé-Casals, A Robust Multiple Feature Approach to Endpoint Detection in Car Environment Based on Advanced Classifiers, IWANN 2005. Lecture Notes in Computer Science, Springer, vol 3512. pp. 850-856.
15. CTIMIT, https://catalog.ldc.upenn.edu/LDC96S30, last accessed September 2019.
16. Dan Ellis's Home Page, Sound Examples for Projects, Columbia University; https://www.ee.columbia.edu/~dpwe/sounds/, last accessed August 2017.
17. Das, R. K., J. Yang, H. Li, Long Range Acoustic Features for Spoofed Speech Detection. Proc. Interspeech, 2019, pp. 1058-1062.
18. Davis, A., S. Nordholm, R. Togneri, Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, 2006, pp. 412-424.
19. Delgado, R., J. Nunez-Gonzalez, Enhancing Confusion Entropy (CEN) for binary and multiclass classification, PLOS ONE, 2019, 14 (1), pp. 1-30.
20. Demuth, H., M. Beale, M. Hagan, Matlab Neural Network Toolbox 6: User's Guide, The MathWorks Inc., 2009.
21. Disken, G., Z. Tufekci, U. Cevik, A robust polynomial regression-based voice activity detector for speaker verification, EURASIP Journal on Audio, Speech, and Music Processing, 2017, vol. 23, pp. 1-16.

22. Dehak, N., P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, V. Hubeika, F. Castaldo, Support Vector Machines and Joint Factor Analysis for Speaker Verification, In: Proceedings of ICASSP, 2009, pp. 4237-4240.
23. Dehak, N., P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-End Factor Analysis for Speaker Verification, *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, vol. 19, No. 4, pp. 788-798.
24. Dwijayanti, S., K. Yamamori, M. Miyoshi, Enhancement of speech dynamics for voice activity detection using DNN, *EURASIP Journal on Audio, Speech, and Music Processing*, 2018:10, pp. 1-15.
25. ELRA-ELDA portal - European Language Resources Association, http://www.elra.info/en/, last accessed September 2019.
26. ETSI, Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms. ETSI ES 202 050 V1.1.5 (2007-01). Annex A.3. Stage 2 – VAD Logic, pp. 42-43.
27. Fazel, A., S. Chakrabartty, An Overview of Statistical Pattern Recognition Techniques for Speaker Verification, in *IEEE Circuits and Systems Magazine*, 2011, vol. 11, No. 2, pp. 62-81.
28. Fawcett, T., An Introduction to ROC analysis, *Pattern Recognition Letters*, vol. 27, No. 8, 2006, pp. 861–874.
29. Feng, Z., J. Feng, F. Dai, The Application of Extreme Learning Machine and Support Vector Machine in Speech Endpoint Detection, International Journal of Control and Automation, 2016, 9(12), pp.191-202.
30. Ferri, C., J. Hernandez-Orallo, R. Modroiu, An experimental comparison of performance measures for classification, *Pattern Recognition Letters*, vol. 30, No.1, 2009, pp. 27–38.
31. Ferrer, L., M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, V. Mitra, A Noise-Robust System for NIST 2012 Speaker Recognition Evaluation, Interspeech, 2013, pp. 1981-1985.
32. FFMTIMIT, https://catalog.ldc.upenn.edu/LDC96S32, last accessed September 2019.
33. Forman, G., I. Cohen, Learning from Little: Comparison of Classifiers Given Little Training. In: Boulicaut JF., Esposito F., Giannotti F., Pedreschi D. (eds) Knowledge Discovery in Databases: PKDD 2004. Lecture Notes in Computer Science, vol 3202. Springer, pp. 161-172.
34. Fukuda, T., O. Ichikawa, M. Nishimura, Long-Term Spectro-Temporal and Static Harmonic Features for Voice Activity Detection, *IEEE Journal of Selected Topics in Signal Processing*, 2010, vol. 4, No. 5, pp. 834-844.
35. Fujioka, K., N. Hayasaka, Y. Miyanaga, N. Yoshida, Noise Reduction of Speech Signals by Running Spectrum Filtering, *Systems and Computers in Japan*, 2006, 37(14), pp. 52-61.
36. Gales, M., S. Young. The Application of Hidden Markov Models in Speech Recognition, Journal Foundations and Trends in Signal Processing, vol. 1, 2008, No 3, pp. 195-304.
37. Ganchev, T., Contemporary Methods for Speech Parameterization, Springer Briefs in Speech Technology, Springer-Verlag, New York, 2011.
38. Ganapathy, S., S. Thomas, H. Hermansky, Temporal envelope compensation for robust phoneme recognition using modulation spectrum, *Jnl. Acoust. Soc. of America*, Vol. 128 (6), 2010, pp. 3769-3780.
39. Ganapathy, S., P. Rajan, H. Hermansky, Multi-layer Perceptron based Speech Activity Detection for Speaker Verification, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011, pp. 321-324.
40. Garcia-Romero, D., C. Espy-Wilson, Analysis of I-vector Length Normalization in Speaker Recognition Systems, Interspeech, 2011, pp. 249-252.
41. Ghaemmaghami, H., B. Baker, R. Vogt, S. Sridharan, Noise robust voice activity detection using features extracted from the time-domain autocorrelation function. In Proceedings of Interspeech, 2010a, pp. 3118-3121.

42. Ghaemmaghami, H., D. Dean, S. Sridharan, I. McCowan, Noise Robust Voice Activity Detection Using Normal Probability Testing and Time-Domain Histogram Analysis, In: Proc. of IEEE ICASSP, 2010b, pp. 4470-4473.
43. Ghosh, P., A. Tsiartas, S. Narayanan, Robust voice activity detection using long-term signal variability, *IEEE Trans. Audio, Speech, Lang. Processing*, 2011, vol. 19, 3, pp. 600–613.
44. Graf, S., T. Herbig, M. Buck, G. Schmidt, Features for voice activity detection: a comparative analysis, *EURASIP Journal on Advances in Signal Processing*, 2015:91, pp.1-15.
45. Gu, L., Zahorian, S.: A new robust algorithm for isolated word endpoint detection, IEEE ICASSP, 2002, vol. 4, pp. 4161-4164.
46. Hain, T. et al., The Development of AMI System for Transcription of Speech in Meetings, Proc. MLMI, 2005, pp. 344-356.
47. Hanilci, C., T. Kinnunen, M. Sahidullah, A. Sizov, Classifiers for Synthetic Speech Detection: A Comparison, Interspeech, 2015, pp. 2057-2061.
48. Hansen, J., T. Hasan, Speaker Recognition by Machines and Humans: A tutorial review, *IEEE Signal Processing Magazine*, November 2015, pp. 74-99.
49. Hasan, T., J. Hansen, Maximum Likelihood Acoustic Factor Analysis Models for Robust Speaker Verification in Noise, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014, vol. 22, No. 2, pp. 381-391.
50. Hegde, R., H. Murthy, V. Gadde, Significance of the Modified Group Delay Feature in Speech Recognition, *IEEE Transactions on ASLP*, vol. 15, 2007, No 1, pp. 190-202.
51. Huang, L. and C. Yang, A Novel Approach to Robust Speech Endpoint Detection in Car Environment, In Proc. of the IEEE ICASSP'2000, pp. 1751-1754.
52. Ichikawa, O., T. Fukuda, and M. Nishimura, Local peak enhancement combined with noise reduction algorithms for robust automatic speech recognition in automobiles, in Proceedings of the IEEE ICASSP, 2008, pp. 4869–4872.
53. Ishizuka, K., T. Nakatani , M. Fujimoto, N. Miyazaki, Noise robust voice activity detection based on periodic to aperiodic component ratio, *Speech Communication*, 52, 2010, pp. 41–60.
54. ISIP, The Institute for Signal and Information Processing, Endpoint detectors, https://www.isip.piconepress.com/projects/speech/software/legacy/signal_detector/index.html, last accessed August 2018.
55. Jain, A., P. Flynn, A. Ross, Handbook of Biometrics, Springer, 2008.
56. Kenny, P., T. Stafylakis, J. Alam, P. Ouellet, M. Kockmann, Joint Factor Analysis for Text-Dependent Speaker Verification, Proc. Odyssey 2014, pp. 123-130.
57. Khoury, E., M. Garland, I-Vectors for Speech Activity Detection, The Speaker and Language Recognition Workshop - Odyssey 2016, pp. 334-339.
58. Kinnunen, T., P. Rajan, A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data, Proceedings of the IEEE ICASSP, 2013, pp. 7229-7233.
59. Kisku, D. (Ed.), Gupta, P. (Ed.), Sing, J. (Ed.). Advances in Biometrics for Secure Human Authentication and Recognition. Boca Raton: CRC Press, 2014.
60. Kitaoka, N., K. Yamamoto, T. Kusamizu, S. Nakagawa, T. Yamada, S. Tsuge, C. Miyajima, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Kuroiwa, K. Takeda, S. Nakamura, Development of VAD evaluation framework CENSREC-1-C and investigation of relationship between VAD and speech recognition performance, ASRU-2007, pp. 607-612.
61. Klapuri, A., Qualitative and quantitative aspects in the design of periodicity estimation algorithms, 10th European Signal Processing Conference, 2000, pp. 1-4.
62. Kola, J., C. Wilson, T. Pruthi, Voice Activity Detection, MERIT BIEN 2011, Final Report, pp. 1-6.

63. Krishnan, S., Padmanabhan, R., Murthy, H.: Robust Voice Activity Detection using Group Delay Functions, In: IEEE Int. Conf. on Industrial Technology, 2006, pp. 2603-2607.
64. Kristjansson, T., S. Deligne, P. Olsen, Voicing features for robust speech detection, Interspeech, 2005, pp. 369-372.
65. Kuncheva, L. Combining Pattern Classifiers: Methods and Algorithms, 2nd Edition, John Wiley & Sons, 2014.
66. Kyriakides, A., C. Pitris, A. Fink, A. Spanias, Isolated word endpoint detection using Time-Frequency Variance Kernels, 2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), Publisher: IEEE, pp. 1041-1045.
67. Larcher, A., K. Lee, B. Ma, H. Li, RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases, Proc. Interspeech, 2012, pp. 123-130.
68. Larcher, A., K. Lee, B. Ma, H. Li, Text-dependent speaker verification: Classifiers, databases and RSR2015, *Speech Communication*, 60 (2014), pp. 56–77.
69. LDC-Linguistic Data Consortium, https://www.ldc.upenn.edu/, last accessed September 2019.
70. LeCun, Y., L. Bottou, G. Orr, K.-R. Müller, Efficient Backprop, Neural Networks, Tricks of the Trade, Lecture Notes in Computer Science LNCS 7700, Springer Verlag, 2012, pp. 9-48.
71. Li, Q., J. Zheng, A. Tsai, Q. Zhou, Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition, *IEEE Transaction on SAP*, vol. 10, No. 3, March 2002, pp. 146-157.
72. Li, J., P. Zhou, X. Jing, Z. Du, Speech Endpoint Detection method based on TEO in Noisy Environment, *Procedia Engineering*, Elsevier Ltd., vol.29, 2012, pp. 2655-2660.
73. Luengo, I., E. Navas, I. Odriozola, I. Saratxaga, I. Hernaez, I. Sainz, D. Erro, Modified LTSE-VAD Algorithm for Applications Requiring Reduced Silence Frame Misclassification, In: Proc. of International Conference on Language Resources and Evaluation (LREC'10), 2010, pp. 1539-1544.
74. Ma,Y., A. Nishihara, Efficient voice activity detection algorithm using long-term spectral flatness measure, *EURASIP J. Audio, Speech, Music Processing*, 2013 (1), pp. 1–18.
75. Macho, D., C. Nadeu, Comparison of Spectral Derivative Parameters for Robust Speech Recognition, Eurospeech 2001, pp. 205-208.
76. Mansour, D., B. Juang, A Family of Distortion Measures Based Upon Projection Operation for Robust Speech Recognition, *IEEE Transaction on ASSP*, 1989, 37, No. 11, pp. 1659-1671.
77. Mak, M., H. Yu, A study of voice activity detection techniques for NIST speaker recognition evaluations, *Computer Speech and Language*, 2014, vol. 28, pp. 295–313.
78. McCowan, I., D. Dean, M. McLaren, R. Vogt, S. Sridharan, The Delta-Phase Spectrum with Application to Voice Activity Detection and Speaker Recognition, *IEEE Trans. Audio, Speech and Language Processing*, 2012, vol. 19, no. 7, pp. 2026–2038.
79. Mesgarani, N., M. Slaney, and S. A. Shamma, Discrimination of speech from non-speech based on multiscale spectro-temporal modulations, *IEEE Trans. Audio, Speech and Language Process.*, 2006, vol. 14(3), pp. 920-930.
80. Misra, H.; Ikbal, S.; Sivadas, S.; Bourlard, H., Multi-resolution Spectral Entropy Feature for Robust ASR, In the Proceedings of the ICASSP, 2005, pp. 253 – 256.
81. Mixer 6 Corpus, https://catalog.ldc.upenn.edu/LDC2013S03, last accessed December 2019.
82. Munteanu, D., S. Toma, Automatic Speaker Verification Experiments Using HMM, In: Proc. of 8th International Conference on Communications, 2010, pp. 107-110.
83. Muralishankar, R., R. Venkatesha Prasad, S. Vijay, H. Shankar, Order Statistics for Voice Activity Detection in VoIP, IEEE International Conference on Communications, 2010, pp. 1-6.
84. Murthy, H., B. Yegnanarayana, Group delay functions and its applications in speech technology, *Sadhana - Indian Academy of Science, Springer Nature*, vol. 36, part 5, 2011, pp. 745-782.

85. Nautsch, A., R. Bamberger, C. Busch, Decision Robustness of Voice Activity Segmentation in unconstrained mobile Speaker Recognition Environment, 2016 International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1-7.
86. NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms, https://ecs.utdallas.edu/loizou/speech/noizeus/, last accessed February 2019.
87. NOISEX, http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html, last accessed March 2014.
88. NIST SRE, NIST Speaker Recognition Evaluation, https://www.nist.gov/itl/iad/mig/speaker-recognition, last accessed September 2019.
89. NTIMIT, https://catalog.ldc.upenn.edu/LDC93S2, last accessed September 2019.
90. Oppenheim, A., R. Schafer, J. Buck, Discrete-Time Signal Processing, Prentice Hall, 2nd ed., 1999.
91. Ouzounov, A., BG-SRDat: A Corpus in Bulgarian Language for Speaker Recognition over Telephone Channels, *Cybernetics and Information Technologies*, 2003, vol. 3, No. 2, pp. 101-108.
92. Ouzounov, A., A Robust Feature for Speech Detection, *Cybernetics and Information Technologies*, vol. 4, No. 2, 2004, pp. 3-14.
93. Ouzounov, A., Robust Features and Neural Network for Noisy Speech Detection, *Cybernetics and Information Technologies*, vol. 6, № 3, 2006, pp.75-84.
94. Ouzounov, A., Further Results on Speaker Identification using Robust Speech Detection and Neural Network, *Cybernetics and Information Technologies*, vol. 9, No. 1, 2009, pp.37-45.
95. Ouzounov, A., Cepstral Features and Text-Dependent Speaker Identification - A Comparative Study, *Cybernetics and Information Technologies*, vol. 10, No.1, 2010, pp.1-12.
96. Ouzounov, A., Noisy Speech Endpoint Detection Using Robust Feature. – In: Proc. of Springer International Publishing Switzerland. V. Cantoni et al., Eds. BIOMET 2014, LNCS 8897, pp. 105-117.
97. Padmanabhan, R., S. Krishnan, H. Murthy, A pattern recognition approach to VAD using modified group delay, in Proc. 14th National conference on Communications, 2008, pp. 432–437.
98. Paint.Net, https://www.getpaint.net/index.html, last accessed January 2020.
99. Padrell, J., D. Macho, C. Nadeu, Robust Speech Activity Detection using LDA applied to FF, In the Proceedings of the ICASSP, 2005, pp. I.557 – I.560.
100. Rabiner, L. and R. W. Schafer, Theory and Application of Digital Speech Processing, Prentice Hall Press, NJ, 2010.
101. Ramirez, J., C. Segura, C. Benitez, A. Torre, A. Rubio, Efficient voice activity detection algorithms using long-term speech information, *Speech Communications*, 2004, vol. 42, pp. 271–287.
102. Ramirez, J., et al. SVM-Based Speech Endpoint Detection Using Contextual Speech Features, Electronics Letters, vol. 42, 2006, No 7, pp. 426-428.
103. Ramirez, J., J. Gorriz, J. Segura, Voice Activity detection. Fundamentals and Speech Recognition System Robustness, In M. Grimm and Kroschel, *Robust speech recognition and Understanding*, 2007, pp. 1-22.
104. Renevey, Ph. and A. Drygajlo, Entropy Based Voice Activity Detection in Very Noisy Conditions, Eurospeech, 2001, pp. 1883-1886
105. Reynolds, D. et al., The 2004 MIT Lincoln laboratory speaker recognition system, Proc. ICASSP, 2005, pp. 177-180.
106. Roach, P., English Phonetics and Phonology: A Practical Course, 4th Ed. Cambridge University Press, 2009.

107. Roy, T., T. Marwala, S. Chakraverty, Precise detection of speech endpoints dynamically: A wavelet convolution based approach, *Communications in Nonlinear Science and Numerical Simulation*, Elsevier Ltd., vol. 67, 2019, pp. 162-175.
108. Sadjadi, S., C. Greenberg, E. Singer, D. Reynolds, L. Mason, J. Hernandez-Cordero, The 2018 NIST Speaker Recognition Evaluation, Interspeech, 2019, pp. 1483-1487.
109. Sahidullah, M., T. Kinnunen, C. Hanilci, A Comparison of Features for Synthetic Speech Detection, Proc. Interspeech, 2015, pp. 2087-2091.
110. Sailor, H., M. Kamble, H. Patil, Auditory Filterbank Learning for Temporal Modulation Features in Replay Spoof Speech Detection. Proc. Interspeech, 2018, pp. 666-670.
111. Scott, D., Scott's Rule, WIREs Computational Statistics, vol. 2, 2010, pp. 497-502.
112. SITW, http://www.speech.sri.com/projects/sitw/, last accessed September 2019.
113. Singer. H., T. Umezaki, F. Itakura, Low Bit Quantization of the Smoothed Group Delay Spectrum for Speech Recognition, Proceedings of ICASSP, 1990, pp. 761-764.
114. Sohn, J., N. Kim, and W. Sung, A statistical model based voice activity detection, *IEEE Signal Processing Letters*, 1999, vol. 6, No. 1, pp. 1-3.
115. Speaker Recognition API (SR-API), Proprietary Developed Software. Internal Reports (unpublished), Fadata, Ltd., 2003.
116. SpEAR Database, Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, http://www.cslu.ogi.edu/nsel/data/SpEAR_lombard.html, last accessed September 2016.
117. SPIDRE, https://catalog.ldc.upenn.edu/LDC94S15, last accessed September 2019.
118. Speech Resources Consortium (SRC), ASJ-Continuous Speech Corpus for Research, http://research.nii.ac.jp/src/en/ASJ-JIPDEC.html, last accessed September 2019.
119. Sturim, D., P. Torres-Carrasquillo, J. Campbell, Corpora for the Evaluation of Robust Speaker Recognition Systems. Proc. Interspeech, 2016, pp. 2776-2780.
120. SWITCHBOARD-1, https://catalog.ldc.upenn.edu/LDC97S62, last accessed September 2019.
121. Theodoridis, S., K. Koutroumbas, An Introduction to Pattern Recognition: A MATLAB Approach, Academic Press, 2010.
122. Tian, X., S. Du, X. Xiao, H. Xu, E. Chng, H. Li, Detecting synthetic speech using long term magnitude and phase information, 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), pp. 611-615.
123. TIMIT, https://catalog.ldc.upenn.edu/LDC93S1, last accessed September 2019.
124. Tistarelli, M., Y. Sun, Cohort-Based Score Normalization. In: Li S., Jain A. (eds) Encyclopedia of Biometrics. Springer, 2014, pp.297-304.
125. Togneri, R., D. Pullella, An Overview of Speaker Identification: Accuracy and Robustness Issues, *IEEE Circuits and Systems Magazine*, vol. 11, No. 2, Second quarter 2011, pp.23-61.
126. Tsiartas, A., T. Chaspari, N. Katsamanis, P. Ghosh, M. Li, M. Van Segbroeck, A. Potamianos, S. Narayanan, Multi-band long-term signal variability features for robust voice activity detection, Interspeech, 2013, pp. 718-722.
127. Tuononen, M., R. Hautamäki, P. Fränti. Automatic voice activity detection in different speech applications. In Proceedings of the 1st international conference on Forensic applications and techniques in telecommunications, information, and multimedia and workshop, e-Forensics, 2008, pp. 12:1-12:6.
128. VAST corpus, https://catalog.ldc.upenn.edu/LDC2019S05, last accessed December 2019.
129. VOICEBOX: Speech Processing Toolbox for MATLAB, last accessed September 2018, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
130. VoxCeleb, http://www.robots.ox.ac.uk/~vgg/data/voxceleb/, last accessed September 2019.
131. Wang, M., Bitzer, D., McAllister, D., Rodman, R., Taylor, J. An Algorithm for V/UV/S Segmentation of Speech. Proceedings of the 2001 International Conference on Speech Processing, pp. 541-546.

132. WaveSurfer, https://sourceforge.net/projects/wavesurfer/, last accessed December 2019.
133. Wei, J.-M., X.-J. Yuan, Q.-H. Hub, S.-Q. Wang, A novel measure for evaluating classifiers, Elsevier, *Expert Systems with Applications*, 2010, vol. 37, pp. 3799–3809.
134. WTIMIT, https://catalog.ldc.upenn.edu/LDC2010S02, last accessed September 2019.
135. Wu, B.-F., K.-C. Wang, Robust Endpoint Detection Algorithm based on the Adaptive Band-Partitioning Spectral Entropy in Adverse Environments, *IEEE Transactions on SAP*, vol. 13, No. 5, 2005, pp. 762-775.
136. Wu, B.-F. and K.-C. Wang, Voice Activity Detection based on Auto-Correlation Function using Wavelet Transform and Teager Energy Operator, in *Computational Linguistics and Chinese Language Processing*, 2006, vol. 11, No. 1, pp. 87-100.
137. Wu, G., Z. Zhu, A. Li, Fuzzy Neural Networks for Speech Endpoint Detection, In: Proc. of 2012 International Conference on Fuzzy Theory and Its Applications, pp. 354-356
138. Yali, C., L. Dongsheng, J. Shuo, N. Xuefen, A Speech Endpoint Detection Algorithm Based on Wavelet Transforms. – In: Proc. of 26th Chinese Control and Decision Conference (CCDC), 2014, pp. 3010-3012.
139. Xunbo L., Z. Chunli, L. Xin, Speech Endpoint Detection Based on Improvement Feature and S-Transform. In: Li K., Fei M., Du D., Yang Z., Yang D. (eds) Intelligent Computing and Internet of Things. ICSEE 2018, IMIOT 2018, Communications in Computer and Information Science, Springer, 2018, vol. 924, pp. 225-235.
140. Yamamoto, K., F. Jabloun, K. Reinhard, A. Kawamura, Robust Endpoint detection for Speech recognition based on Discriminative Feature Extraction, Proc. ICASSP, 2006, pp. 805-808.
141. Yamamoto, H., K. Okabe, and T. Koshinaka, Robust i-vector extraction tightly coupled with voice activity detection using deep neural networks, in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017, pp. 600– 604.
142. Ying, D., Y. Yan, J. Dang, F. Soong, Voice Activity Detection Based on an Unsupervised Learning Framework, IEEE Trans. on ASLP, 2011, vol. 19, no. 8, pp. 2624– 2633.
143. Yoo, I.-C., H. Lim, D. Yook, Formant-based Robust Voice Activity Detection, *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, No. 12, 2015, pp. 2238-2245.
144. Zhang, Z., S. Furui, Noisy Speech Recognition Based on Robust End-Point Detection and Model Adaptation, Proceedings of IEEE ICASSP, Vol. 1, 2005, pp. 441-444.
145. Zhang, H., H. Hu, An Endpoint Detection Algorithm Based on MFCC And Spectral Entropy Using BP NN, IEEE 2010 2nd International Conference on Signal Processing Systems (ICSPS), pp. 509-513.
146. Zhang, T., H. Huang, L. He, M. Lech, A Robust Speech Endpoint Detection Algorithm Based on Wavelet Packet and Energy Entropy, In: Proc. of 3rd International Conference on Computer Science and Network Technology, 2013, pp. 1050-1054.
147. Zhang, Y., K. Wang, B. Yan, Speech endpoint detection algorithm with low signal-to-noise based on improved conventional spectral entropy, 12th World Congress on Intelligent Control and Automation (WCICA), 2016, pp. 3307-3311.
148. Zhu, D., K. Paliwal, Product of Power Spectrum and Group Delay Function for Speech Recognition, In Proceedings of 2004 IEEE International Conference on ASSP, pp. 125-128.
149. Zou, Y., W. Zheng, W. Shi, H. Liu, Improved Voice Activity Detection Based on Support Vector Machine with High Separable Speech Feature Vectors, 2014, 19th International Conference on Digital Signal Processing, Hong Kong, pp. 763-767.
150. Тилков, Д., Т. Бояджиев, Българска фонетика, София, Наука и Изкуство, 1977.
151. Узунов, А., Изследване на спектъра на групово закъснение при зашумени с адитивен шум говорни сигнали, сп. Автоматика и Информатика, 1993, 11/12, стр. 14-16.