Б Ъ Л Г А Р С К А А К А Д Е М И Я Н А Н А У К И Т Е
ИНСТИТУТ ПО ИНФОРМАЦИОННИ И КОМУНИКАЦИОННИ ТЕХНОЛОГИИ

Атанас Петров Узунов

ДЕТЕКЦИЯ НА ГОВОР В СИСТЕМИ ЗА РАЗПОЗНАВАНЕ НА ДИКТОРИ

А В Т О Р Е Ф Е Р А Т

НА ДИСЕРТАЦИЯ

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

Обща характеристика на дисертационния труд

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 съдбовно 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 разглеждани в дисертационния труд.

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.

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 и 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.

Методология на изследването

Препоръчаната методология се основава на методи и подходи от следните области (но без ограничения само върху тях):

- линейна алгебра – линейни трансформации, векторни пространства и др.;
- цифрова обработка на сигнали – корелационен анализ, спектрален анализ и др.;
- разпознаване на образи – невронни мрежи, марковски модели и др.

Съдържание на дисертацията

Дисертационния труд се състои от речник на термините, увод, пет глави, резюме на получените резултати, публикации по дисертацията и цитирания и библиография. Първа глава е озаглавена – *“Детекция на говор. Аналитичен преглед“*, втора глава – *“Дефиниране на признаци за детекция на говор използващи свойствата на САКФ и СГЗ“*, трета глава *–“Алгоритми за определяне на гранични точки при зависима от текста верификация на диктори. Експериментално изследване“*, четвърта – *“Алгоритми за детекция на говор при независима от текста идентификация на диктори. Експериментално изследване“* и пета *–“BG-SRDat – Корпус с говорни данни записани по телефонен канал и предназначен за разпознаване на диктори”.*
Основното съдържание е изложено на 164 страници. Включени са 48 фигури и 27 таблици. Библиографията включва общо 151 цитирани източника.

Глава 1. Детекция на говор. Аналитичен преглед.

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.).

Детекцията на говор е означена в англоезичната литература с различни термини като 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.

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 [Nautsch et al., 2016].

1.2. VAD-алгоритми

Детекторите на говор (VAD-алгоритми) 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 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]. При същия тип верификация на диктори и NIST 2016 SRE е използвана невронна мрежа с дълбочинно обучение [Yamamoto et al., 2017].

1.2.1. Признаци използвани при VAD- и ED-алгоритмите

В текста са описани 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].

1.2.2. Класификатори използвани при VAD- и ED-алгоритмите

В текста са разгледани основните класификатори намерили приложение при VAD-алгоритмите: модел с Гаусови смеси, метод на опорните вектори и метод с i-вектори.

1.2.3. VAD-алгоритми в системи за разпознаване на диктори

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, които са специално адаптирани за NIST 2010 SRE. Особеността при тези тестове за верификация на диктори е качеството на записите. В една немалка част от тях SNR е около 5 dB. Използвани са две системи за верификация на диктори. При първата е приложен GMM-SVM подхода [Campbell et al., 2006a], а при втората метода с i-вектори [Dehak et al., 2011]. 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, 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 (Equal Error Rate).

1.2.3.3. VAD-алгоритъм на базата на MLP

Предложения алгоритъм [Ganapathy et al., 2011] се базира на апостериорните вероятности на фонемите в английския език, получени на изходите на многослоен перцептрон. 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. 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 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.

1.3. Определяне на гранични точки на кратки фрази

1.3.1. Увод

Алгоритмите за ОГТ включват 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 [Zhang et al., 2013], [Zhang et al., 2016], времево-честотни параметри [Kyriakides et al., 2011], МЕЛ-кепстър [Cao et al., 2017], уейвлети [Yali et al., 2014] и др. На втория етап, най-често се използва или краен автомат [Chung et al., 2014] или класификатор - с невронни мрежи [Wu et al., 2012], скрити Марковски модели (Hidden Markov Models - HMM) [Zhang et al., 2005], метод на опорните вектори (Support Vector Machines - SVM) [Feng et al., 2016] и др.

1.3.2. Алгоритми за определяне на гранични точки

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 EOT 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 EOT 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 EOT.

1.4. Заключение

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 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 при промяна в условията на средата.

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

В работата 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 тя се използва. В работата е прието SACF да се изчислява директно чрез амплитудния спектър със спектрална резолюция аналогична на тази на преобразуването на Фурие. Ако $|X(k)|$ е амплитудния спектър на говорния сигнал, получен чрез FFT за текущия сегмент, то изместената оценка на автокорелационната функция $R_A(l)$ съгласно [Klapuri, 2000] има вида

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

където $l=0,...,L$ и $L=K/2-1$; K е размерът на FFT и L е броят на отместванията (lags).

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. В този случай ортогоналния полиномиален коефициент от първи ред описва наклона на кепстралната траектория за даден времеви сегмент. Тези коефициенти са известни под името ‘делта кепстрални коефициенти’ или просто делта кепстър.

Друга интерпретация на делта-кепстъра е предложена в [Fukuda et al., 2010] където той се разглежда като последователност получена на изхода на линеен филтър с крайна импулсна характеристика. Съгласно [Fukuda et al., 2010] този филтър има предавателна характеристика $H(z)$ от вида

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

В работата е предложено производната на спектралната автокорелационна функция да се разглежда като изходен сигнал на филтъра в (2.5). Тази производна е означена като Delta Spectral Autocorrelation Function (DSACF) и се изчислява като се използват стойностите на SACF в границите само на текущия сегмент. В текста SACF е означена като $R(n,l)$ независимо от това как е изчислена - чрез амплитудния или чрез мощностния спектър. Вида на спектъра е указан допълнително в текста.

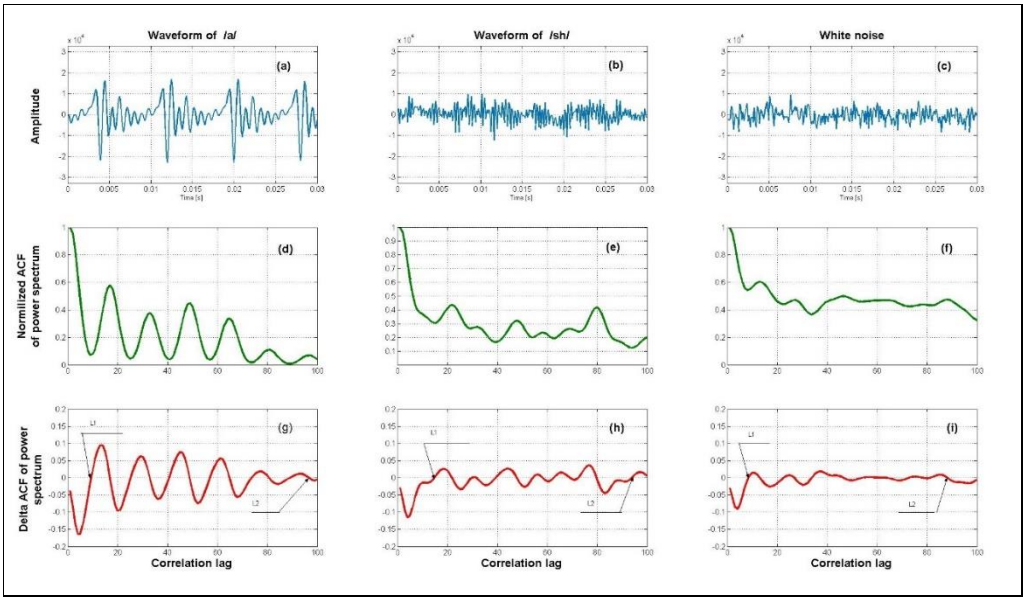
DSACF $\Delta R(n,l)$ за *n-я* сегмент се изчислява чрез SACF $R(n,l)$ съгласно

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

където $l=0,...,L$ е броят изместванията (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, нормализираните им SACF и съответните DSACF (Q=3) до лаг $L=100$.

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 2nd FFT в [Wang et al., 2001], 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 (2nd FFT spectrums) и да се установи ефекта от прилагането на делта филтъра (2.5) върху спектралната автокорелационна функция.



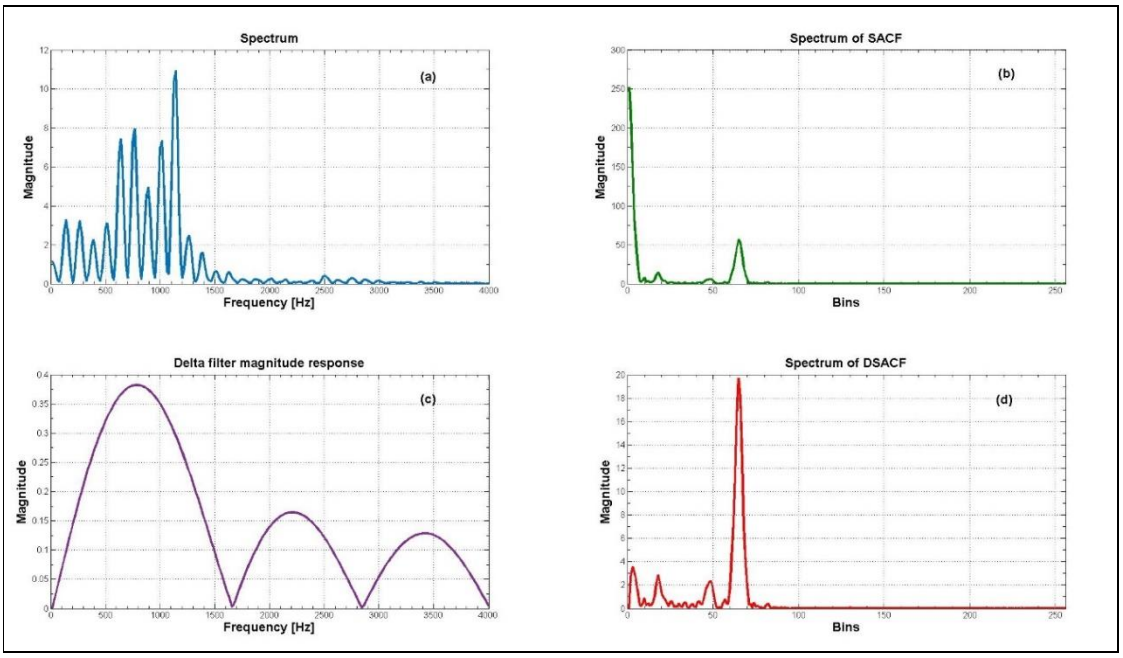
Фиг. 2.3 Нормирани SACF и съответните DSACF за (a) звучна фонема */a/*, (b) шумова фонема */ш/* и (c) бял шум.

На фиг. 2.4 са показани съответно: на 2.4 (a) - амплитудния спектър на част от фонемата *'a'* (времевия ред е показан на фиг. 2.3 (a) - честота на дискретизация 8 kHz), на 2.4. (b)- амплитудния спектър на SACF $\left|S_R(\Omega)\right|$, на 2.4 (c) - амплитудната честотна характеристика на КИХ филтъра (2.5) $\left|H(\Omega)\right|$ и на 2.4 (d) - амплитудния спектър на DSACF $\left|S_{\Delta R}(\Omega)\right|$. 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 е получена съгласно

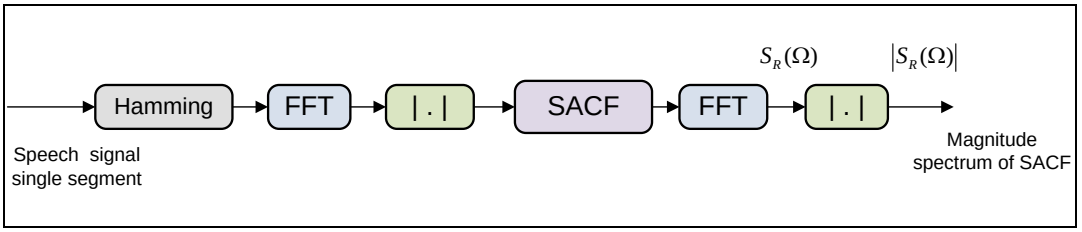
$$R(l)=\begin{cases}\dfrac{1}{K/2-|l|}\displaystyle\sum_{k=0}^{K/2-1-l}\left|X(k)\right|.\left|X(k+l)\right|; & l\geq 0;\ l\leq L;\\ R(-l) & ;l<0\end{cases}, \qquad (2.7)$$

където K е брой точки на FFT, $L=K/4$ и $\left|X(.)\right|$ е амплитудният спектър на говорния сигнал за даден сегмент.

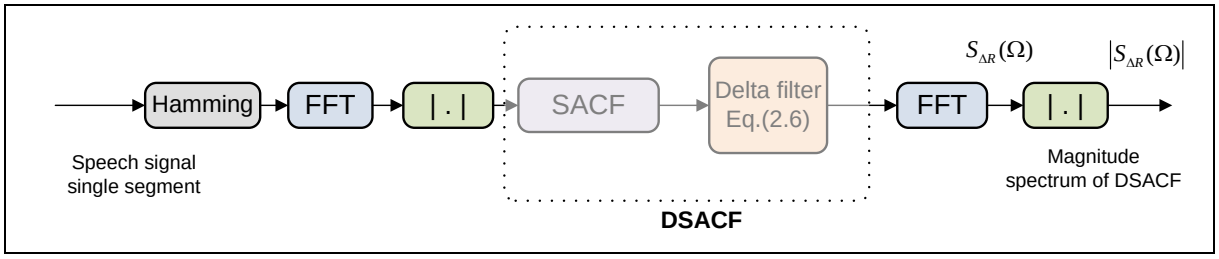
Блоковите схеми на алгоритмите за изчисляване на $\left|S_R(\Omega)\right|$ и $\left|S_{\Delta R}(\Omega)\right|$ са показани на фиг. 2.5 и фиг. 2.6.



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



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 стойностите на честотите са отразени в Таблица 2.1. С f_L и f_H са означени честотите

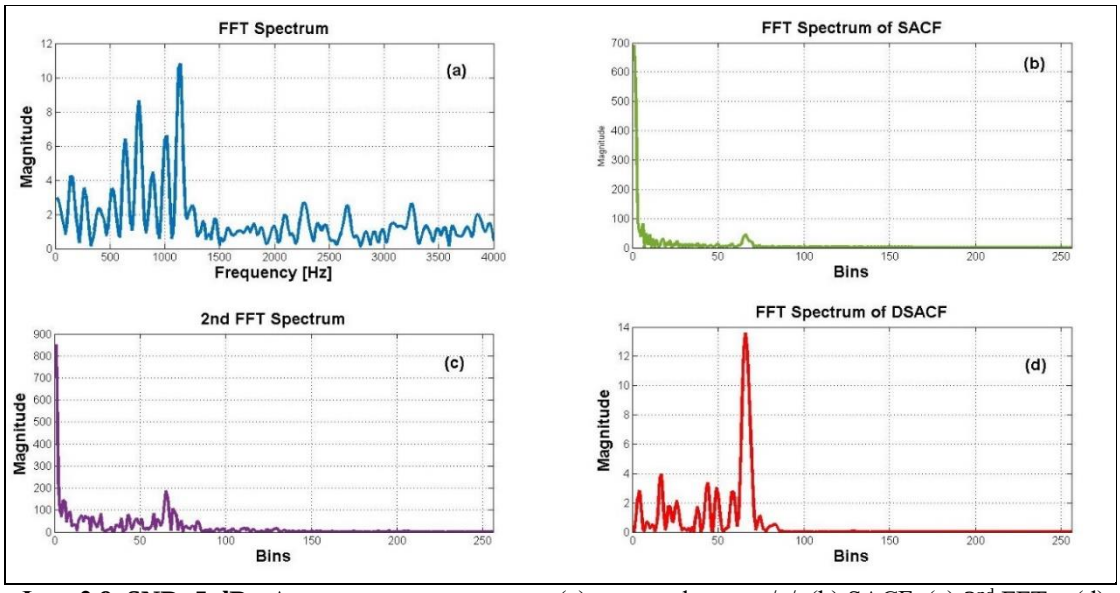
на срязване, f_0 е централната честота и B е лентата на пропускане на съответните лентови филтри.

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).

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-А. 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 за 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. 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) алгоритъма с параметър *J=3* [Ramirez et al., 2004]. Така получената изгладена версия на $\Delta R(n,l)$ е означена като $\Delta R^s(n,l)$

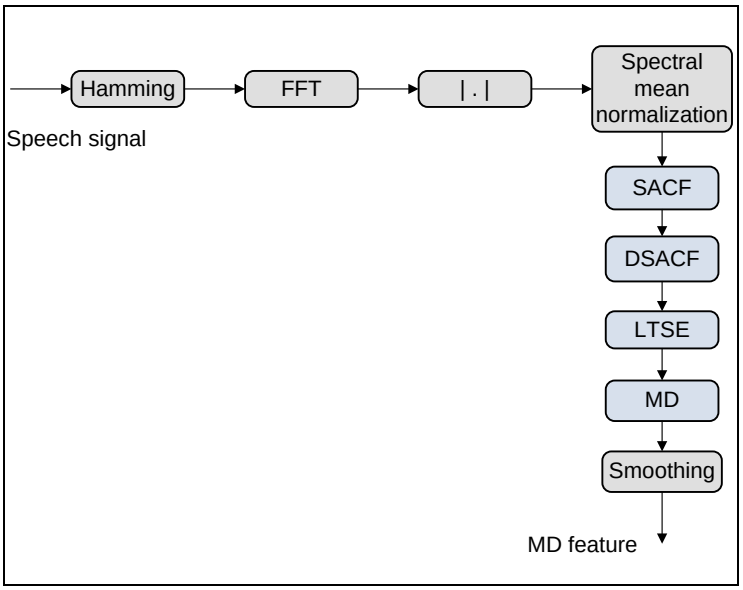
$$\Delta R^s(n,l) = \max\left\{\Delta R(n+j,l)\right\}_{j=-J}^{j=+J}. \quad (2.16)$$

- изчислява се MD признака $m_d(n)$

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

- изглажда се контура на $m_d(n)$ 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.



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

2.1.4.2. Базов среден делта признак

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

- за претегления с тегловна функция на Хеминг говорен сигнал се изчислява амплитудния спектър $\left|X(k)\right|$ чрез FFT с размер *K*;
- определя се амплитудния спектър $\left|\widehat{X}(n,k)\right|$ 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 $\Delta R^s(n,l)$

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

- 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

В тази подточка е разгледан Спектъра на Групово Закъснение (СГЗ) - (Group Delay Spectrum - GDS) [Murthy et al., 2011] и е извършен анализ на неговото изменение 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) съчетаващо Модифицирания СГЗ (МСГЗ) - Modified GDS – (MGDS) [Hegde et al., 2007] и MD признака, разгледан в т.2.1.4.

2.2.2. Спектър на групово закъснение

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

2.2.4. Group Delay Mean-Delta признак

MGDS $\tau_m(\omega)$ 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 GDS и да се постигне допълнително усилване на съответните пикове в делта спектралната автокорелационна функция. Предложени са две модификации на GDMD-признака. Предложените GDMD-признаци за $n^{-я}$ сегмент се изчисляват 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

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

  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

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

В1. 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 контура на m_{gd-lin} чрез филтър с изместваща се средна стойност;

В2. Трети етап-2– определяне на log-GDMD контура и изглаждане:

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

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

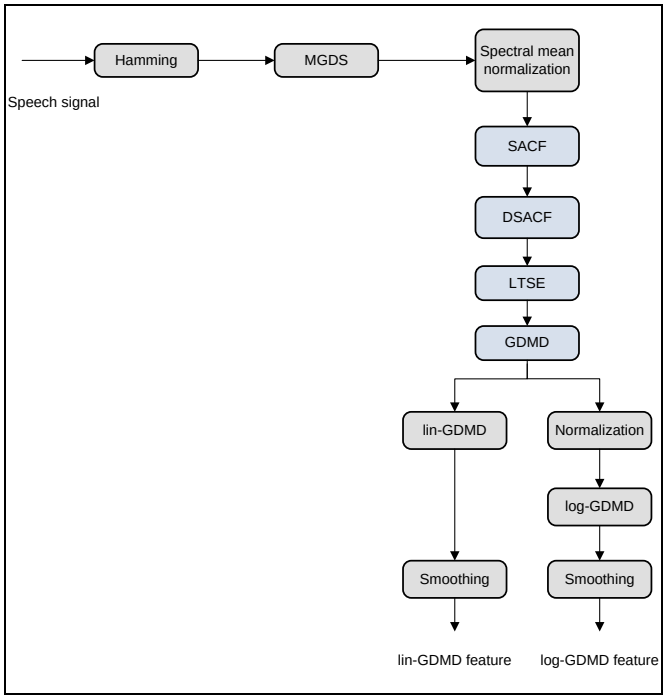
  където $m_{gd}^{\min} = \min_n\{m_{gd}(n)\}$.

- определяне на log-GDMD съгласно

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

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

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.



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

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. Във втората част на главата е извършен качествен анализ на изменението на GDS при зашумени с адитивен шум говорни сигнали. 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.

Глава 3. Алгоритми за определяне на гранични точки при зависима от текста верификация на диктори. Експериментално изследване

3.1. Увод

В тази 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.

Оценката на ефективността е реализирана в три етапа. На първия етап за даден признак е определено Евклидовото разстояние между неговите два Z-нормализирани времеви контура - получени съответно от чистата и зашумената версия на тестовата фраза [Chen et al., 2005]. 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].

На втория етап е оценена точността на различните детектори 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].

На третия етап е оценено влиянието на алгоритмите за определяне на граничните точки върху точността на разпознаване 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 избрани от корпуса BG-SRDat. 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].

3.2. Референтни признаци

Описани са признаците изброени по-горе в текста.

3.3. Анализ на Z-нормализирани времеви контури
3.4. Алгоритми за определяне на гранични 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, който включва алгоритъм за изчисляване на адаптивни прагове и детерминиран краен автомат. 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.

3.4.2. Алгоритъм за определяне на адаптивни прагове

За да се намали грешката при детекция на говор чрез използване на фиксирани прагове тук е предложен алгоритъм с адаптивни прагове, 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 общия 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 $l_{\min}=\min\limits_{l}\{Q_M\}$ и $l_{\max}=\max\limits_{l}\{Q_M\}$. 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 ма 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:

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

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

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.

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.

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 текущия файл с аудио данни не постъпва към следващите нива на системата за разпознаване. Грешка се генерира при следните условия:

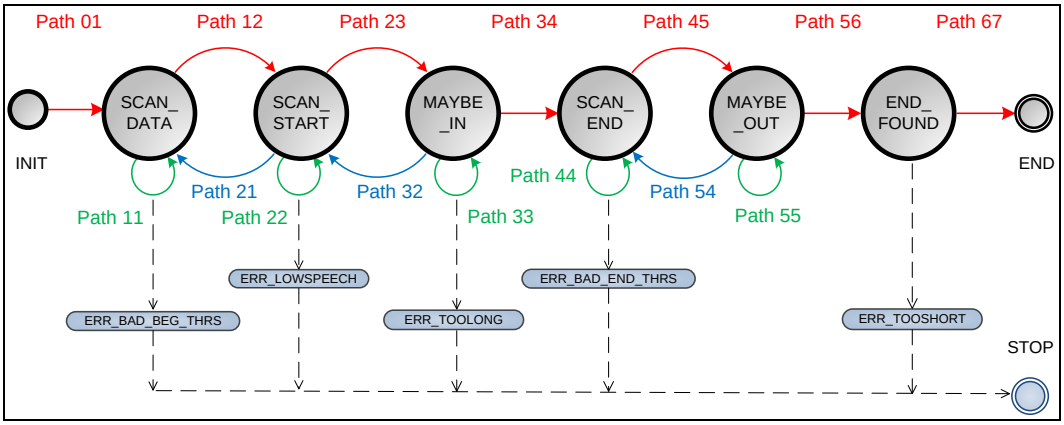
- когато произнасянето завършва извън аудио файла (не се изпълняват условията за край на говорното съобщение) - error ERR_TOOLONG;
- когато SNR е много ниско - error ERR_LOWSPEECH;
- когато праговете не позволяват да бъдат определени началните или крайните точки (лошо установени прагове) - errors ERR_BAD_BEG_THRS, 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.

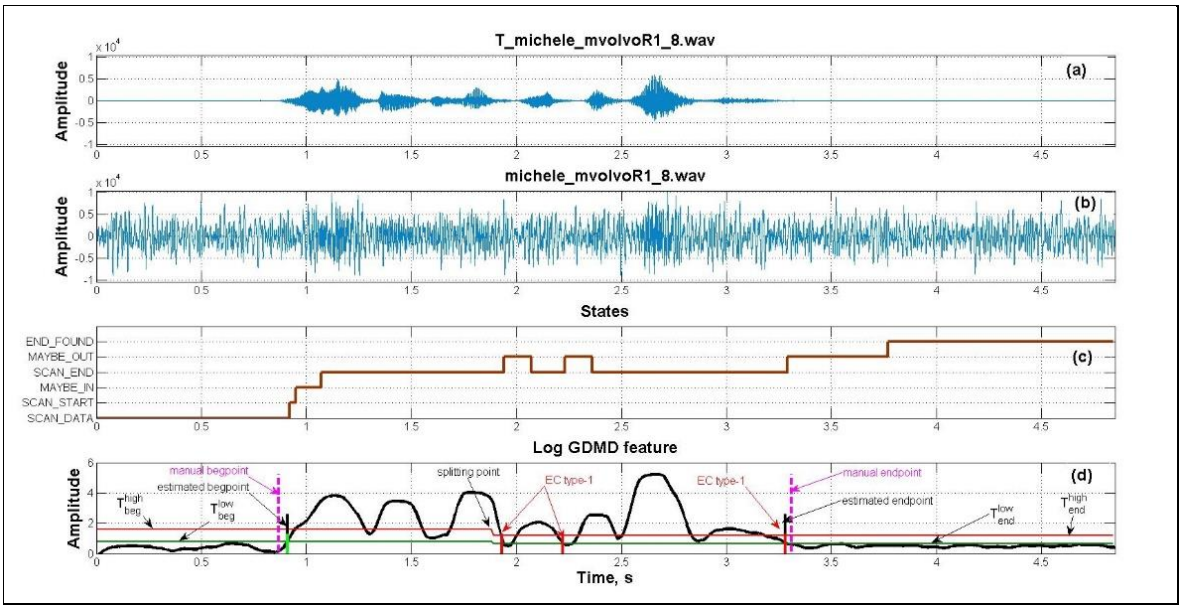
Логиката на предложения краен автомат е показана на фиг. 3.12. Параметрите $T_{\text{SCAN\_START}}$, $T_{\text{MAYBE\_IN}}$, $T1_{\text{SCAN\_END}}$, $T2_{\text{SCAN\_END}}$ и $T_{\text{MAYBE\_OUT}}$ са таймери показващи колко време автоматът е в дадено състояние (state timers). Въведени са времеви константи *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 крайна гранична точка, от които чрез логически правила се избира финалната крайна точка.

Като илюстрация на резултатите от предложения алгоритъм (с адаптивни прагове) на фиг. 3.13 е показан контура на признака log-GDMD за зашумен сигнал от корпуса SpEAR. 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_{beg}^{low}, T_{beg}^{high} и крайните точки T_{end}^{low}, T_{end}^{high}.



Фиг. 3.12. Диаграма на логиката на крайния автомат.

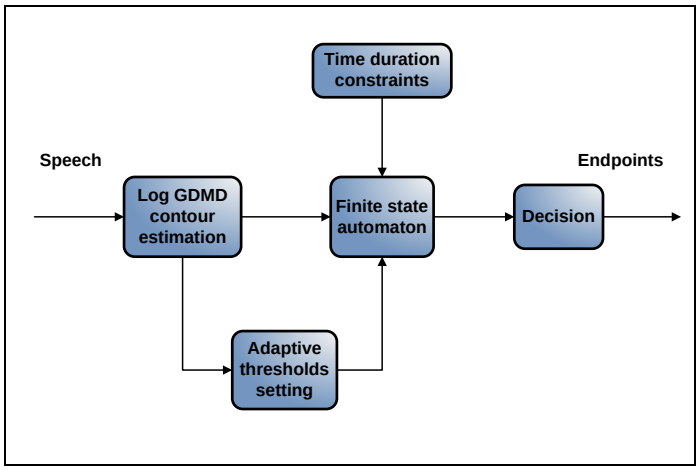


Фиг. 3.13. Запис от корпуса SpEAR: (a) чист сигнал; (b) зашумена версия; (c) времева диаграма на преходите; (d) контур на log-GDMD с маркирани някои характерни детайли от изпълнението на предложения алгоритъм във времето.

3.5. Детектори на гранични точки

3.5.1. Детектор GDMD-E

Този детектор предложен като 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). Блок схемата на GDMD-E детектора е показана на фиг. 3.15.



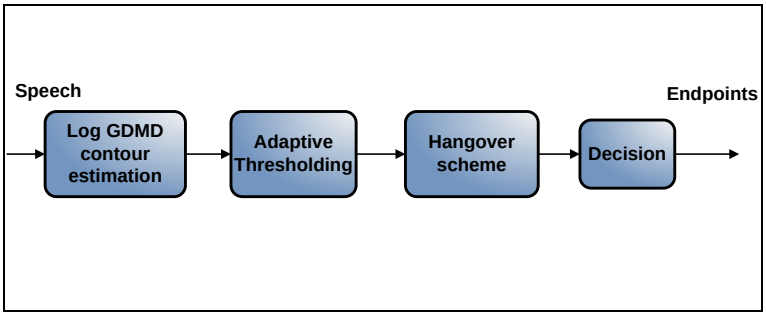
Фиг. 3.15. Блок схема на детектор GDMD-E.

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.

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 признака. 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

3.6.1. Говорни данни

Говорните данни, използвани в експериментите са избрани от два корпуса - BG-SRDat [Ouzounov, 2003] и TIDIGITS [Dan Ellis, 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. От корпуса BG-SRDat са избрани записи, чиито дължина 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). От корпуса TIDIGITS, който е на английски език, са избрани записи съдържащи последователности от цифри. Данните включват 84 файла, събрани от 6 диктора. 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.

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

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

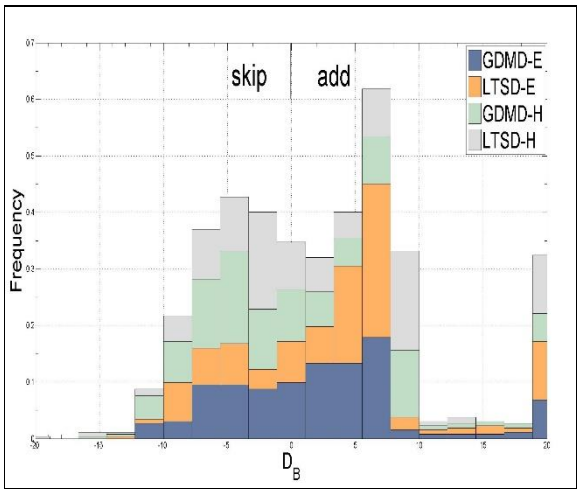
където $M_E(s)$ е ръчно определената крайна гранична точка; $ED_E(s)$ е крайната гранична точка, определена чрез съответния алгоритъм.

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 признак 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 вероятността в проценти сегментните разлики да бъдат в указаните диапазони (rate of the distribution). Прието е, че значими при детекция диапазони на изменение на сегментните разлики са [-10; 10] и [-5; 5]. Абсолютните стойности на сегментните разлики са означени в таблиците като $|D_B|$ и $|D_E|$, а с $\bar{D}$ е означена тяхната средна стойност. В Таблица 3.4 са показани резултатите за корпус BG-SRDat съответно с адаптивни прагове (adapt2thr) и логаритмична скала (log scale).

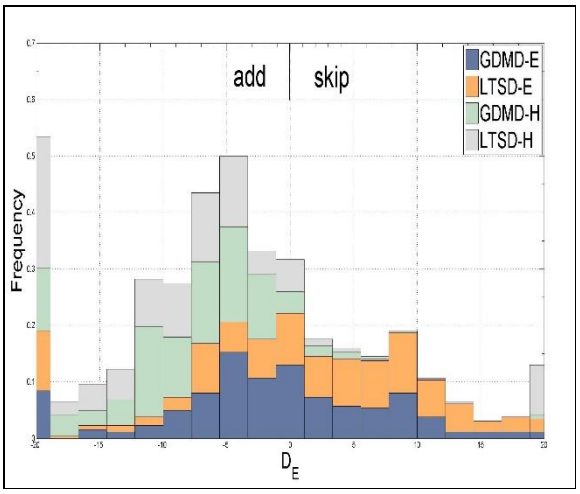
Таблица 3.4. Вероятност в проценти

Speech Corpus		**BG-SRDat**					
No.	**Features & adapt2thr**	**$\|D_B\|$**		**$\|D_E\|$**		**$\bar{D}$**	
		≤ 5	**≤ 10**	**≤ 5**	**≤10**	**≤ 5**	**≤ 10**
1	**log-MTE-E**	56.10	71.37	**55.72**	77.09	55.91	74.23
2	**log-EE-E**	49.61	65.26	36.64	58.01	43.12	61.64
3	**log-MD-E**	**60.30**	80.91	54.96	74.42	**57.63**	77.67
4	**log-GDMD-E**	54.96	**87.02**	51.90	**78.24**	53.43	**82.63**
5	**LTSD-E**	41.60	84.35	37.02	67.17	39.31	75.76
6	**log-GDMD-H**	47.32	**87.02**	35.11	61.06	41.22	74.04
7	**LTSD-H**	45.80	85.11	24.42	46.56	35.11	65.83

Като цяло (от гледна точка на средните стойности в приетите диапазони) най-добри резултати са получени при параметри в логаритмичната скала, съчетани с алгоритъма за адаптивни прагове. Избрани 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.



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



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

На хистограмите представени във фигури 3.18 - 3.19 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. 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.

3.6.4. Зависима от текста верификация на диктори

В тази част от работата всеки един от изброените 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. Използвани са два различни подхода при верификация на диктори DTW и HMMs. Тестовете са реализиран с кратки фрази на български език избрани от корпуса BG-SRDat.

3.6.4.1. Предварителна обработка

Като параметрични представяне се използва Мел-кепстъра (Mel-Frequency Cepstral Coefficients - MFCC) с размерност 14 получен чрез 24 Мел-филтри с еднаква площ. За всеки файл се прилага кепстрална нормализация [Ganchev, 2011].

3.6.4.2. Верификация на диктори чрез DTW

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 различно от анализираните диктори. 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].

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 модела (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

Известно е, че при корпуси с малък брой данни и освен това записани в реална среда 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). 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. 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.

Таблица 3.10. HMM – Грешки при верификация на диктори

No.	Endpoint detector	FRR, %	FAR, %	HTER, %	95%CI
1	**Manual**	15.63	1.21	8.42	±0.0134
2	**log-GDMD-E**	18.45	0.98	9.71	±0.0143
3	**LTSD-E**	22.25	1.20	11.72	±0.0153
4	**log-GDMD-H**	18.45	1.02	9.73	±0.0143
5	**LTSD-H**	22.53	1.04	11.78	±0.0154

3.7. Заключение

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; **Второ** –точността на детекция при използване на краен автомат с адаптивни прагове винаги превъзхожда, при един и същи признак, тази получена чрез hangover алгоритъма и **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.

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 детекцията е реализирана чрез бинарна класификация. Независимо от широкото използване на алгоритми за VAD засега не съществува универсален алгоритъм, който да работи надеждно в реална среда (real-world environment).

В работата е реализиран сравнителен експериментален анализ на ефективността на предложените в Глава 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 многослоен перцептрон (МСП).

Експерименталните изследвания са реализирани с два различни алгоритма за детекция на говор –означени в текста като VAD-1 и VAD-2. Причината за разработването на два детектора е, че в Гл. 2 са предложени два вида признаци – във векторна форма (VAD-1) и в скаларна форма (VAD-2).

Във VAD-1 като класификатор е използван МСП и бинарното решение се получава чрез прагов алгоритъм анализиращ стойностите, получени в изходните неврони на перцептрона. Работата на VAD-2 се основава на анализ на контури и бинарното решение се получава чрез сравняването на техните стойности с прагове (подобно на подходите разгледани в предишната глава). Към системата за разпознаване на диктори се подават само говорните сегменти маркирани като такива от съответния детектор [Kitaoka et al., 2007].

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.

На втория етап е оценено влиянието на различните VAD-алгоритми върху точността на разпознаване в система за независима от текста идентификация на диктори, в която като класификатор се използва МСП. Тук тестовете са реализирани с фрази на български език избрани от корпуса BG-SRDat.

4.2. Референтни признаци

Като референтни признаци 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 признаци са съответно получените чрез алгоритъма на Sohn [Sohn et al., 1999] (разгледан в 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].

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. Често използваните грешки са описани в [Davis et al., 2006]. Те са: Front-End Clipping (FEC), Mid-speech Clipping (MSC), OVER (over hang**),** Noise Detected as Speech (NDS), Back-End Clipping (BEC), Front-end adding (FEA) и Speech Detected as Noise (SDN). Точността при детекция е - вярно разпознати говорни сегменти или Speech Hit Rate (SHR) и вярно разпознати не-говорни сегменти или Non-speech Hit Rate (NHR).

4.4. Оценка на точността при детекция и грешката от 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). Най-често обаче се въвеждат 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], [Delgado et al., 2019].

4.4.1. ROC-анализ

4.4.2. Матрица на грешките

4.5. Идентификация на диктори независимо от текста

4.6. Детектор на говор – VAD-1

4.6.1. Използвани признаци

В детектора на говор са използвани четири референтни признака и два предложени от автора. За всеки признак е реализиран отделен детектор. Признаците са: предложени от автора - BMD и MMD в т. 2.1.4.2-3 и референтни MBSE в т. 4.2.2, FF в т. 4.2.3, RSD в т. 4.2.3 и IW-MFCC в т. 4.2.4.

4.6.2. Многослоен перцептрон

Структурата на използвания в детектора МСП 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].

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.

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.

4.6.5. Определяне на точността при детекция

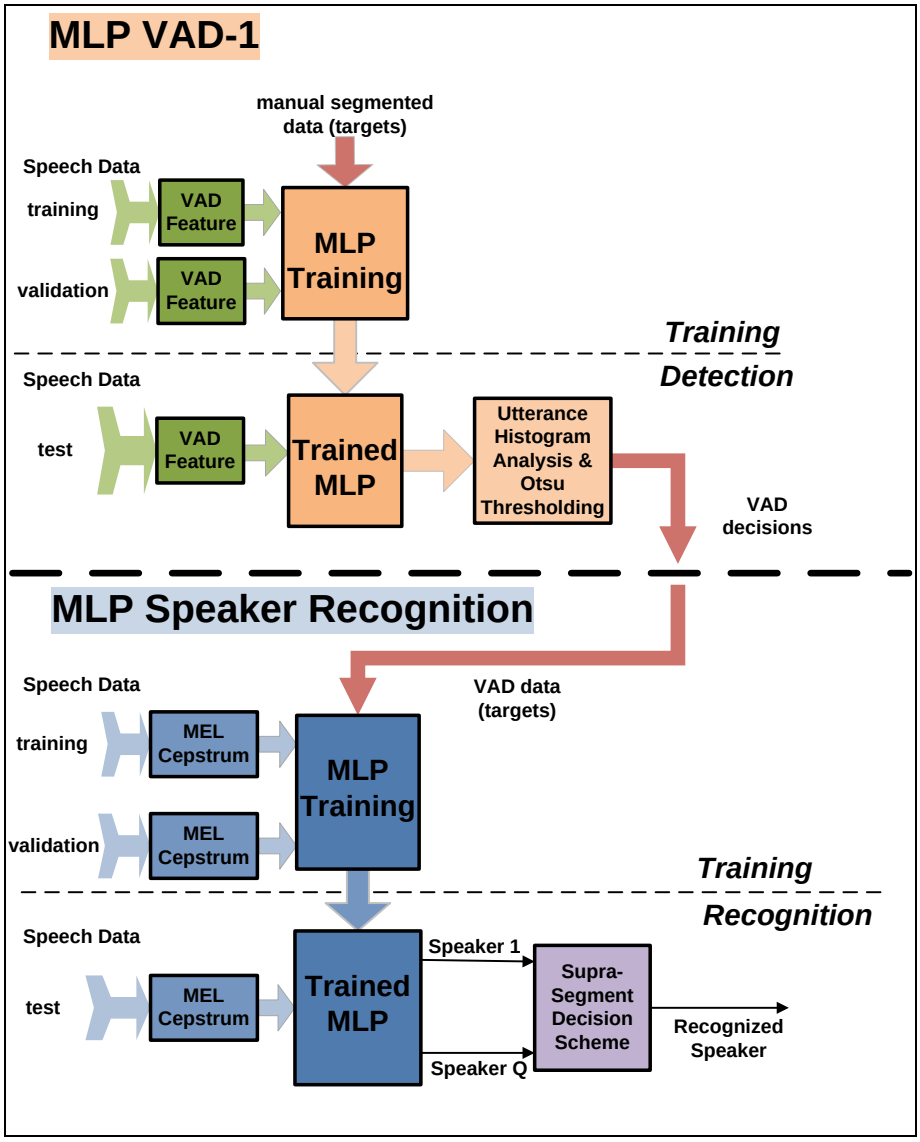
За всички използвани параметри в таблица 4.1 са показани стойностите на AUC, F-measure и грешките при детекция в проценти. Те са изчислени като претеглени средни стойности на резултатите от всички тествани файлове – общо 270 файла.

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, МСП-класификатор и модул за вземане на решение чрез анализ на последователност от супра-сегменти (supra-segment decision scheme). Блоковата схема на системата за разпознаване на диктори чрез МСП включваща детайли за модула VAD-1 е показана на фиг. 4.3.

4.7.1. Предварителна обработка

Модула за предварителната обработка включва два под-модула – VAD-1 (описан в т. 4.6) и за определяне на кепстралните параметри - 14 коефициента на МЕЛ-кепстъра получени чрез 24-лентови филтри с еднаква площ. Обработват се само говорните сегменти маркирани от VAD-детектора.



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

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

4.7.2. Многослоен перцептрон

Броят на дикторите, които ще бъдат разпознавани е 12 и структурата на използвания в работата МСП 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]. 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].

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 данните за обучение са използвани еднакъв брой говорни сегменти за всеки клас. Броят на говорните сегменти от един файл е 1300. За всеки диктор са използвани 2 файла или 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).

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. Експерименталните резултати

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. В таблицата се наблюдава съгласуваност между грешката и общата ентропия за първите два и последните два признака. 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 признака MMD.

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

4.8. Детектор на говор – VAD-2

Предложения в работата алгоритъм VAD-2 включва 2 етапа – изчисляване 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: предложен от автора - log-GDMD в т.2.2.4 и референтните - оценката $\Gamma(m)$ получена чрез алгоритъма на Sohn в т. 1.2.3.1.1; SAE – получен чрез алгоритъма на Wu в т. 4.2.1 и LTSD описан в т. 3.2.4.

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 признака log-GDMD (отбелязано е в съответната таблица).

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 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].

4.8.4. Определяне на точността при детекция.

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. Съгласно резултатите показани в таблица 4.12 за данни от корпуса на Dan Ellis стойността на AUC е максимална при признака log-GDMD.

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.

4.9.1. Експериментални резултати.

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 4.11. 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

Таблица 4.13. Корпус BG-SRDat - грешка 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

4.10. Заключение

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.
- **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 обаче 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.

ГЛАВА 5. BG-SRDat – Корпус с говорни данни записани по телефонен канал и предназначен за разпознаване на диктори

5.1. Увод

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

5.3. Кратко описание на известни корпуси

5.4. Описание на BG-SRDat

Съобразно вида на говорния материал корпуса BG-SRDat може да се разглежда като съставен от 5 модула (SD1, 2,…,5), които 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). Подобно фонемно съдържание затруднява разпознаването на диктори в телефонен канал.
- SD3 (BG) - съдържа запис на прочетени текстове (със средна продължителност от около 80 секунди) избрани от различни вестници. Реализирани са 14 записи от 10 диктора чрез стационарен и мобилен телефон. Текстове са избрани така, че да се получи в някаква степен лексикално разнообразие.
- SD4 (BG) – запис на разговори с максимална продължителност от около 7 минути. Реализирани са 4 записа от 4 диктора чрез мобилен телефон и VoIP. Разговорите са на свободна тема като в някои фрагменти едновременно разговарят и двамата диктори.
- SD5 (EN) - съдържа запис на кратка фраза на английски език. Реализирани са 150 записа от 9 диктора чрез стационарен телефон.

При описание на корпуса ще бъде обърнато внимание на следните атрибути: 1) вида на говорния материал; 2) брой сесии и периода между тях; 3) условия, при които са реализирани записите и 4) описание на файловете.

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

За всеки файл е създадена структура от метаданни съдържаща информация от вида: ID на диктора; вид говорни данни; място, от където се провежда разговора; вид шум и др. Понастоящем само метаданни за записи получени чрез мобилни телефони постепенно се въвежда в MySQL база данни.

5.5. Приложение на BG-SRDat

Корпуса е използван за изследвания в областта на верификация на диктори чрез фиксирани фрази, 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, които са получени само от мобилни устройства.

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 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).
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.

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 отпечатаните научни публикации по дисертацията:

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.26-29, 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. 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

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. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. Dan Ellis's Home Page, Sound Examples for Projects, Columbia University; https://www.ee.columbia.edu/~dpwe/sounds/, last accessed August 2017.
11. 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.
12. Delgado, R., J. Nunez-Gonzalez, Enhancing Confusion Entropy (CEN) for binary and multiclass classification, PLOS ONE, 2019, 14 (1), pp. 1-30.
13. Demuth, H., M. Beale, M. Hagan, Matlab Neural Network Toolbox 6: User's Guide, The MathWorks Inc., 2009.
14. 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.
15. 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.
16. 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.
17. Fawcett, T., An Introduction to ROC analysis, *Pattern Recognition Letters*, vol. 27, No. 8, 2006, pp. 861–874.
18. 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.
19. 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.
20. 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.
21. Ganchev, T., Contemporary Methods for Speech Parameterization, Springer Briefs in Speech Technology, Springer-Verlag, New York, 2011.

22. 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.
23. 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.
24. Garcia-Romero, D., C. Espy-Wilson, Analysis of I-vector Length Normalization in Speaker Recognition Systems, Interspeech, 2011, pp. 249-252.
25. 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.
26. 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.
27. 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.
28. Gu, L., Zahorian, S.: A new robust algorithm for isolated word endpoint detection, IEEE ICASSP, 2002, vol. 4, pp. 4161-4164.
29. Hain, T. et al., The Development of AMI System for Transcription of Speech in Meetings, Proc. MLMI, 2005, pp. 344-356.
30. 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.
31. 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.
32. 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.
33. Jain, A., P. Flynn, A. Ross, Handbook of Biometrics, Springer, 2008.
34. 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.
35. Kisku, D. (Ed.), Gupta, P. (Ed.), Sing, J. (Ed.). Advances in Biometrics for Secure Human Authentication and Recognition. Boca Raton: CRC Press, 2014.
36. 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.
37. Klapuri, A., Qualitative and quantitative aspects in the design of periodicity estimation algorithms, 10th European Signal Processing Conference, 2000, pp. 1-4.
38. Krishnan, S., Padmanabhan, R., Murthy, H.: Robust Voice Activity Detection using Group Delay Functions, In: IEEE International Conference on Industrial Technology, 2006, pp. 2603-2607.
39. Kuncheva, L. Combining Pattern Classifiers: Methods and Algorithms, 2nd Edition, John Wiley & Sons, 2014.
40. 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.
41. 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.
42. 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.
43. 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.
44. Macho, D., C. Nadeu, Comparison of Spectral Derivative Parameters for Robust Speech Recognition, Eurospeech 2001, pp. 205-208.
45. 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.

46. 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.
47. 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.
48. 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.
49. Munteanu, D., S. Toma, Automatic Speaker Verification Experiments Using HMM, In: Proc. of 8th International Conference on Communications, 2010, pp. 107-110.
50. 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.
51. 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.
52. NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms, https://ecs.utdallas.edu/loizou/speech/noizeus/, last accessed February 2019.
53. NIST SRE, NIST Speaker Recognition Evaluation, https://www.nist.gov/itl/iad/mig/speaker-recognition, last accessed September 2019.
54. Oppenheim, A., R. Schafer, J. Buck, Discrete-Time Signal Processing, Prentice Hall, 2nd ed., 1999.
55. 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.
56. 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.
57. Rabiner, L. and R. W. Schafer, Theory and Application of Digital Speech Processing, Prentice Hall Press, NJ, 2010.
58. 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.
59. 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.
60. Renevey, Ph. and A. Drygajlo, Entropy Based Voice Activity Detection in Very Noisy Conditions, Eurospeech, 2001, pp. 1883-1886
61. Reynolds, D. et al., The 2004 MIT Lincoln laboratory speaker recognition system, Proc. ICASSP, 2005, pp. 177-180.
62. Roach, P., English Phonetics and Phonology: A Practical Course, 4th Ed. Cambridge University Press, 2009.
63. Scott, D., Scott's Rule, WIREs Computational Statistics, vol. 2, 2010, pp. 497-502.
64. 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.
65. 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.
66. 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.
67. Theodoridis, S., K. Koutroumbas, An Introduction to Pattern Recognition: A MATLAB Approach, Academic Press, 2010.
68. 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.
69. VOICEBOX: Speech Processing Toolbox for MATLAB, last accessed September 2018, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
70. 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.
71. 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.

72. 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.
73. 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.
74. Wu, G., et al., Fuzzy Neural Networks for Speech Endpoint Detection, In: Proc. of 2012 International Conference on Fuzzy Theory and Its Applications, pp. 354-356
75. Yali, C. et al. A Speech Endpoint Detection Algorithm Based on Wavelet Transforms. – In: Proc. of 26th Chinese Control and Decision Conference (CCDC), 2014, pp. 3010-3012.
76. 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.
77. 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.
78. 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.
79. 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.
80. 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.
81. 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.
82. 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.
83. 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.
84. Тилков, Д., Т. Бояджиев, Българска фонетика, София, Наука и Изкуство, 1977.