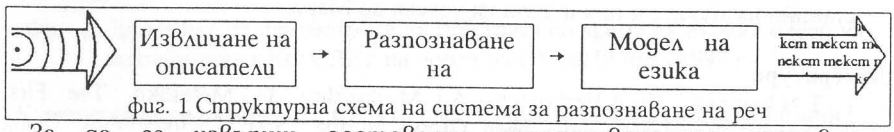
About descriptor estimation for speech recognition

inj. Dragomir D. Nikolov, TU of Varna. assoc. prof. Ph.D. Jordan N. Kolev, TU of Varna.

One of the fundamental and probably the most important questions in speech recognition problem is the optimal choice of speech descriptors. The method proposed by the authors is a modification of the well known Mel Frequency Cepstral Coefficients (MFCC) which is applied for discrete in time speech and discrete in frequency speech spectra. The main feature of the method is an additional scaling of the frequency axis. Additionally a new sequence of steps is used for estimation of the descriptors. The feature vector for a given vowel can be distinguished from the one for other vowels by more than 53% of the coefficients. Applied to the Bulgarian language the method proposed makes possible complete distinction of all six vowels. The coefficient estimation is adapted for the purpose of the following processing by neural networks. A study of noise influence is also presented. A comparison of the results arrived at using this method and others using different methods is performed. A dictor dependent speech recognition system has been developed for testing the descriptor estimation.

I. Въведение в проблема

Структурната схема на всяка система за разпознаване на реч (СРР) е съставена от два основни блока и един допълнителен (фиг. 1). Входният речеви сигнал постъпва за обработка в блока за извличане на описатели. В този етап от алгоритъма се определят параметрите, характеризиращи сигнала. Извлечените от сигнала характеристики се използват от втория блок за разпознаване. На изхода на този блок се получава разпознатата структурна единица, изграждаща речевия сигнал. Така разпознатите градивни елементи се подлагат на обработка в третия блок. В него са заложени семантични и лингвистични правила на съответния език. При системи, извършващи разпознаване на отделни (цели) думи, този блок е излишен.



фиг. 1 Структурна схема на система за разпознаване на реч

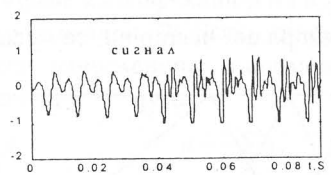
За да се извърши достоверно разпознаване на градивната структурна единица е необходимо да се определят най-важните характеристики на речевия сигнал. Предложеният от авторите алгоритъм е модификация на метода, изчисляващ изместените по Мел-скалата кепстрални коефициенти (ИМКК[1]). С тези коефициенти е постигната грешка под 0,7% [1] при разпознаване на свързани цифри.

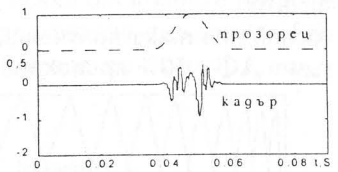
Настоящият метод е разработен при реализацията на цялостна СРР. Детайлите от алгоритъма са тествани и доразвивани на базата на изпробването в реална СРР. Той е адаптиран за работа с невронни мрежи (НМ), което не е ограничение за други апарати.

[1] в западната литература известен като MFCC

II. Описание на предлагания метод.

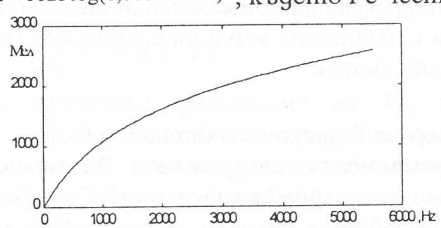
Речевият сигнал с помощта на прозорец се накъсва на отделни кадри. Умножението на входния сигнал с използваната прозоречна функция намалява краевите ефекти и смущенията. В резултат от тази функция се получава стабилно поведение на цялата система и значително подобряване на резултатите (вж. табл. 2).



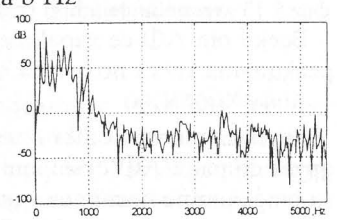


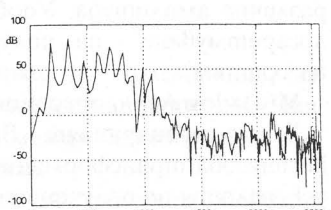
Същността на алгоритъма е използването на изкривяващата скала на Мел. Тя се изчислява чрез формулата:

$m = 1125\log(0{,}0016 \cdot f + 1)$, където f е честотата в Hz



фиг3. Мел-скала





фиг. 4 спектър след изкривяване по Мел скалата

Кривата отразява по-голямата чувствителност на човешкото ухо към ниските честоти. Съществуват няколко подобни скали, но тази е ефективно математическо решение, позволяващо лесно изчисляване. Важна особеност на алгоритъма е, че стойностите на скалата се използват в целия диапазон, за разлика от класическия вариант, в който тя е линейна за честоти под 1 kHz.

За да се намали излишъка от информация се използват система от лентови филтри (ЛФ). Те са разположени линейно по Мел-скалата. Те определят амплитудите на няколко честотни ленти. Тези амплитуди са еквивалент на системата от ЛФ в човешкото ухо, възбуждащи нервните клетки. В дадения метод амплитудите се получават чрез конволюция на ЛФ със спектъра на кадъра. Централните честоти на филтрите се разполагат линейно през **k** единици по скалата на Мел .

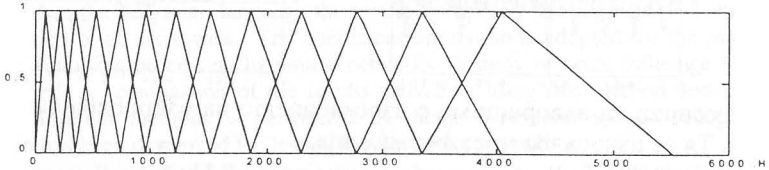
$f_c(p) = m(k \cdot p)$, p - номер на лентовия филтър

Спектърът на всеки кадър (**X(n)**) се изчислява с помощта на бързо преобразуване на Фурие (БПФ). Поради дискретността на спектъра, в дадения алгоритъм централните честоти на ЛФ се коригират. Корекцията се състои в избор на най-близката честота **f'** съществуваща в дискретния спектър, получен след БПФ.

Където **f'** е тази честота, за която се минимизира разликата:

$|f'_p(n) - f_c(p)|$

Въз основа на така получените централни честоти, се построяват триъгълни ЛФ с 50% припокриване.



фиг.5 13 лентови филтри (триъгълни), линейно разположени по Мел-скалата

Всеки от ЛФ се характеризира с честотна характеристика $H_p(n)$, а реакцията му се получава чрез конволюция:

$A(p) = X(n) * H_p(n)$

Речевият сигнал има голяма енергия в нискочестотната област. От друга страна ЛФ сумират определен честотен диапазон. В резултат, за отделните честотни диапазони, се получават амплитуди с коренно различна амплитуда. Удобни за обработка сигнали се получават след логаритмуване на получените амплитуди (част от класическия алгоритъм):

$M(p)=\log A(p)$, операцията е част от хомоморфната обработка.

В модифицираният вариант алгоритъма използва дискретна косинусова трансформация (ДКТ). Идеята е породена от [2]. Целта е декорелация на получените коефициенти и компресия на спектралната информация. Това позволява последващо статистическо моделиране, използвайки диагонални ковариационни матрици.

$C(p)=dct(M(p))$, където **dct** е дискретна косинусова трансформация.

В резултат се получават коефициенти с много добри показатели (вж. табл.1), удобни за обучение на системи на основата на СММ. Голямата разлика между отделните коефициенти не позволява директното им използване при НМ. Изследванията показаха, че ефективността на метода се повишава значително, ако първия коефициент се раздели на 10.

$C'(1)=C(1)/10$

По този начин коефициентите стават съизмерими и се съкращава неколкократно процеса на обучение на НМ.

III. Експериментални изследвания и заключения.

• Сравнение на метода с други алгоритми.

Степента на разделяне на отделните класове може да се оцени от между класовата ковариационната матрица (КМ). Аналогично се оценява и степента на вътрешно разпръскване във всеки клас, чрез изчисляване на вътрешно класовата КМ $\mathbf{W_j}$ [3].

Разглеждаме класификатор на **n** думи, всяка от които е представена с **k** свои реализации. За всяка отделна реализация е съставен вектор-стълб от изчислените параметри, обозначен като $\mathbf{f_{ij}}$ за $\mathbf{i}^{\text{-mama}}$ реализация на $\mathbf{j}^{\text{-mama}}$ дума.

$$\mu_j = \frac{1}{k}\sum_{i=1}^{k} f_{ij} \quad , \qquad \mu = \frac{1}{n}\sum_{j=1}^{n} \mu_j \quad ,$$ където:

μ_j - средноаритметичен вектор за думата **j**

μ - среден вектор за всички думи.

Ковариационната матрица за $\mathbf{j}^{\text{-mama}}$ дума се дефинира като:

$$W_j = \frac{1}{k}\sum_{i=1}^{k}(f_{ij} - \mu_j)(f_{ij} - \mu_j)^T$$

На практика различите на $\mathbf{W_j}$ за отделните класове не се различават съществено, поради което вътрешно-класовата дисперсия може да се изрази чрез осреднената матрица $W=\{W_j\}$. Междукласовата КМ се дефинира от средните вектори за всички думи:

$$B = \frac{1}{n}\sum_{i=1}^{n}(\mu_j - \mu)(\mu_j - \mu)^T$$

Най-често използваният метод за оценка на големината на получените дисперсии е отношението на детерминантите им. При резултат $|\mathbf{B}|/|\mathbf{W}| > 1$, средното разделяне между класовете е по-голямо отколкото техните вътрешни дисперсии.

В таблица 1 са показани резултатите за различни методи, пресметнати по описания метод [3]. Анализът е извършен върху 600 кадъра от мъжки глас.

таблица 1

Метод	забележка	$\lvert\mathbf{B}\rvert/\lvert\mathbf{W}\rvert$
КЛП	коефициенти на линейно предсказване (LPC)	$-8.4834*10^{-140}$
кепстър		$-7.1789*10^{-146}$
ВЛП	възприемчиво линейно предсказване (PLPC)	$1.3478*10^{-141}$
ИМКК	класически ленти, без нормализация (MFCC)	$3.3130*10^{-140}$
ИМКК	класически ленти	$-5.0259*10^{-140}$
ИМКК	ДКТ, класически ленти	$-2.6790*10^{-140}$
ИМКК	ОДКТ, модифицирани ленти	$4.5620*10^{-140}$
ИМКК(М)	ДКТ, модифицирани ленти, без нормализация	$5.0849*10^{-141}$
ИМКК(М)	ДКТ, модифицирани ленти, с нормализация	$8.5926*10^{-138}$

• Предварителна обработка.

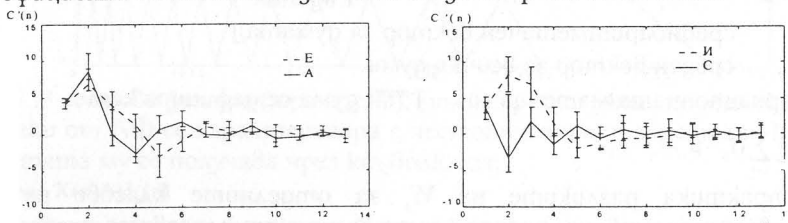
Резултатите от различни прозоречни функции са показани в таблица 2. Много добри резултати се получават, ако се приложи Хеммингов прозорец.

таблица 2

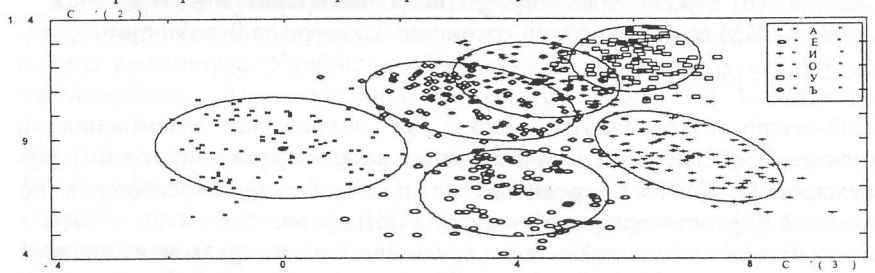
Прозоречна функция	|B|/|W|	Прозоречна функция	|B|/|W|
правоъгълна	$1.0000*10^{-142}$	Хенингова (hanning)	$5.8255*10^{-140}$
триъгълна	$1.3013*10^{-140}$	Хемингова (hamming)	$8.5926*10^{-138}$

• Разпределение на параметрите.

На фиг.6 са показани средните коефициенти C'(n), заедно с техните максимални отклонения. На фигура 7 е показано разпределението на шестте гласни в пространството дефинирано от втори и трети коефициент. Показаните данни са на диктор от мъжки пол.



фиг.6 сравнение на параметрите на звуците "А" и "Е", "И" и "С"

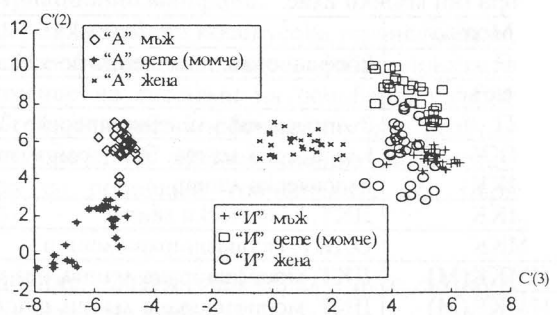


фиг. 7 Разпределение на гласните по $2^{ри}$ и $3^{ти}$ коефициент

•Дикторозависимост.

На фигура 8 са показани измененията на коефициентите в зависимост от пола на диктора, както и възрастта.

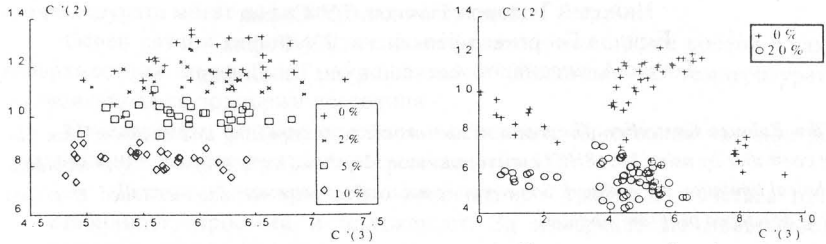
Изследването на параметрите показа, че те значително се влияят от честотата на основния тон.



фиг.8 Дикторо-зависимост на параметрите

•Влияние на шума

На фигура 9 са показани промените в коефициентите, настъпващи



а) промяна на гласната "У" б) промяна на всичките гласни

фиг. 9 Промяна на параметрите при промяна на отношението сигнал/шум.

при смесването на сигнала с бял шум. При ниско съотношение сигнал/шум (С/Ш) се намалява вътрешно класовата дисперсия на коефициентите. Увеличаването на съотношението С/Ш води до преместване на целия клас от описатели. При голямо ниво С/Ш коефициентите вече описват характеристиките на самия шум.

• Практическа реализация.

Крайният разпознавател е реализиран на базата на трислойна НМ. Дикторо-зависимата СРР е обучавана за разпознаване на 12 думи, като за всяка дума са подавани по 18 образци. Разпознаваемоста на тази система е 96.88% за непознати образци (по 32 на дума).

• Ресурси.

За изчисляването на 13 модифицирани ИМКК са необходими 17426 операции с плаваща запетая. Резултата е получен в среда MATLAB, като ЛФ са изчислени предварително и записани като коефициенти. Тази стойност е получена за кадър с дължина 256 отчета и прилагане на Хеммингов прозорец.

Процедурата може да се ускори: ако Хемминговият прозорец е изчислен предварително; ако не се извършва конволюция в целия спектър, а само за онези честоти на лентовите филтри, които са различни от 0.

• Различия на модифицирания алгоритъм от класическия.

♦ Всички лентови филтри се изчисляват спрямо скалата на Мел.

♦ Подравняване със съществуващи в дискретния спектър честоти

♦ Дискретна Косинусова Трансформация

♦ Нормализация на коефициентите

[1] IEEE Spectrum, December 1997, p 39-47

[2] IEEE Signal Processing, vol. 13 No 5. September 1996, p 45-57

[3] Richard A. Haddad, Thomas W. Parsons "Digital Signal Processing theory,applications, and hardware"