STUDIA INFORMATICA 2003
Volume 24 Number 3 (55)

Adam DUSTOR
Politechnika Śląska, Instytut Elektroniki

WERYFIKACJA MÓWCY I JEJ ZASTOSOWANIA W INFORMATYCE

Streszczenie. W pracy przedstawiono podstawowe wiadomości z zakresu technologii weryfikacji i identyfikacji tożsamości mówcy na podstawie jego głosu. Opisano przykładowe zastosowania systemów weryfikacji i identyfikacji mówców w informatyce. Zaprezentowano badania dokładności rozpoznawania na przykładzie systemu stworzonego w środowisku Matlab. Do testów wykorzystano zasób mowy ROBOT, zawierający wypowiedzi mówców o dobrej jakości, nagranych w warunkach biurowych. Zbadano wpływ modelu mówcy i rodzaju parametrów ekstrahowanych z sygnału mowy na proces rozpoznawania.

Słowa kluczowe: weryfikacja mówcy, identyfikacja mówcy, biometria, zasób mowy, ekstrakcja parametrów.

SPEAKER VERIFICATION AND ITS APPLICATIONS IN INFORMATION SCIENCE

Summary. This paper presents fundamentals of speaker identification and verification. Applications of this technology in information science are also included. Obtained performance of speaker verification and identification system constructed in Matlab environment is shown. All tests were done on the basis of Polish speech corpus ROBOT containing utterances of good quality recorded in an office environment. The influence of speaker model and extracted speech parameters on recognition results is also examined.

Keywords: speaker verification, speaker identification, biometrics, speech corpus, feature extraction.

1. Wprowadzenie

Technologie związane z przetwarzaniem mowy ludzkiej stanowią obszar intensywnych badań naukowych od wielu dziesięcioleci. Szczególne zainteresowanie naukowców tą dziedziną nauki związane jest z potencjalnie ogromnymi perspektywami stojącymi przed tego typu technologiami. Zagadnienia związane z przetwarzaniem mowy ludzkiej stanowią bardzo szeroką dyscyplinę naukową, w której wykorzystuje się efekty badań interdyscyplinarnych poczynając od fizjologii słuchu poprzez lingwistykę, fonetykę i fonologię, aż po bardzo zaawansowane algorytmy cyfrowego przetwarzania sygnałów i elementy sztucznej inteligencji.

Problematykę przetwarzania sygnału mowy można podzielić na zagadnienia związane z kompresją sygnału mowy, jej syntezą, rozpoznawaniem i identyfikacją mówcy czy też języka, którym dana osoba się posługuje. Niestety, pomimo upływu wielu lat do tej pory udało się rozwiązać w zadowalający sposób tylko problem kompresji i syntezy sygnału mowy. Kompresja znalazła szerokie zastosowanie w transmisji mowy poprzez sieci komputerowe zarówno lokalne, jak i Internet (VoIP) oraz w telefonii komórkowej. Istnieje co najmniej kilkadziesiąt różnych algorytmów pozwalających ze standardowego strumienia danych, wynoszącego 64 kbit/s zejść do nawet kilkuset bit/s. W zadowalający sposób rozwiązano również problem syntezy mowy z tekstu (*text to speech synthesis*), o czym można się przekonać odwiedzając stronę www firmy Scansoft [12] bądź też instalując jeden z wielu dostępnych na rynku syntezerów mowy [14]. Niestety problem nadal stanowi rozpoznawanie mowy i identyfikacja osoby na podstawie jego głosu. Jedną z przyczyn odpowiedzialnych za ten stan rzeczy jest z pewnością obszerność zagadnienia, jego stopień skomplikowania jak i to, że poprawne rozpoznawanie mowy a w mniejszym stopniu i mówcy wymaga pewnych elementów sztucznej inteligencji oraz ogromnych mocy obliczeniowych komputerów. Choć postęp w tej dziedzinie jest widoczny, to jednak nadal komputer rozumie mowę w sposób bardzo ograniczony. Z ciekawostek należy wymienić, iż firma Microsoft od wielu lat intensywnie rozwija technologię rozpoznawania mowy i jej syntezy udostępniając interfejs programisty SAPI [9,13] dla osób tworzących własne aplikacje głosowe. Istnieją wszakże profesjonalne programy, które już obecnie można z powodzeniem stosować w codziennym użytkowaniu komputera [6,12]. Pozwalają one co prawda na rozumienie mowy ciągłej, lecz niestety nie jest to jeszcze mowa spontaniczna i stanowią głównie nieocenioną pomoc dla osób niepełnosprawnych w dostępie i korzystaniu z komputera oraz dla tych, którzy dużo i często piszą. Trochę inaczej jest w przypadku identyfikacji i weryfikacji mówcy, gdyż już obecnie maszyna osiąga porównywalne a często nawet lepsze rezultaty od człowieka. Mimo to szersze zastosowanie tej technologii wymaga dalszych badań i udoskonaleń.

Zasadniczym celem pracy jest przybliżenie tematyki związanej z rozpoznawaniem osób na podstawie ich głosu oraz przedstawienie potencjalnych zastosowań tej technologii w informatyce, jak również własnych rezultatów badań w tej dziedzinie przeprowadzonych w środowisku Matlab na podstawie zasobu mowy języka polskiego ROBOT [1].

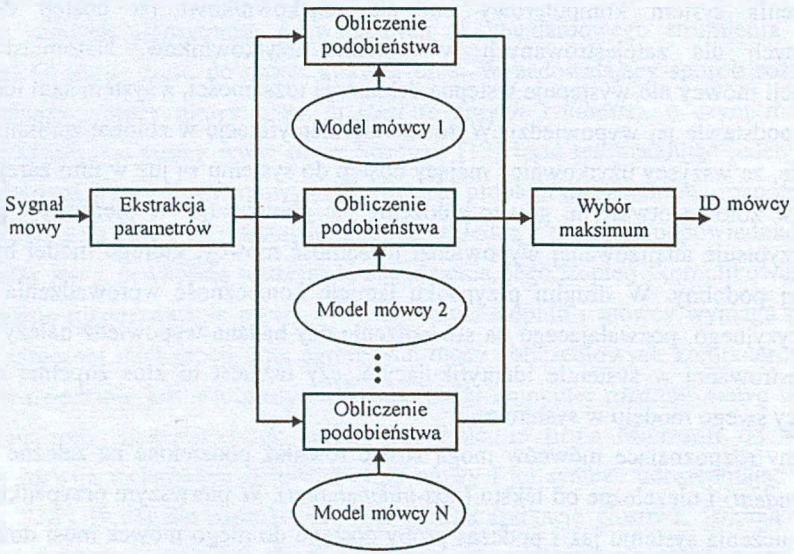
2. Rozpoznawanie mówcy

Rozpoznawanie osób na podstawie ich głosów jest blisko spokrewnione z problematyką rozpoznawania mowy. Cechą odróżniającą te dwa zagadnienia jest fakt, że w rozpoznawaniu mowy istotne jest wydobycie zawartości lingwistycznej (znaczenia, sensu) z analizowanej wypowiedzi, podczas gdy w rozpoznawaniu mówców wyekstrahowanie cech sygnału mowy specyficznych dla danego mówcy. Problem rozpoznawania obejmuje identyfikację i weryfikację. W procesie weryfikacji podejmowana jest decyzja czy badana wypowiedź została wypowiedziana przez mówcę o deklarowanej tożsamości, czego efektem jest potwierdzenie bądź odrzucenie deklarowanej przez użytkownika tożsamości. W przypadku potwierdzenia system komputerowy pozwala użytkownikowi na dostęp do miejsc zastrzeżonych dla zarejestrowanych w systemie użytkowników. Natomiast podczas identyfikacji mówcy nie występuje wstępna deklaracja tożsamości, a system sam identyfikuje osobę na podstawie jej wypowiedzi. Wyróżnia się identyfikację w zbiorze zamkniętym, gdy zakłada się, że wszyscy użytkownicy mający dostęp do systemu są już w nim zarejestrowani bądź też w zbiorze otwartym, gdy to założenie nie obowiązuje. W pierwszym przypadku system przypisuje analizowanej wypowiedzi tożsamość mówcy, którego model był do niej najbardziej podobny. W drugim przypadku istnieje konieczność wprowadzenia pewnego progu decyzyjnego, pozwalającego na stwierdzenie czy badana wypowiedź należy do osoby już zarejestrowanej w systemie identyfikującym, czy też jest to ktoś zupełnie nowy, nie posiadający swego modelu w systemie.

Systemy rozpoznające mówców mogą zostać również podzielone na zależne od tekstu (*text-dependent*) i niezależne od tekstu (*text-independent*). W pierwszym przypadku zarówno w trakcie uczenia systemu jak i podczas próby dostępu do niego mówca musi dostarczyć tę samą wypowiedź. W drugim przypadku wypowiedź mówcy wykorzystywana do stworzenia jego wzorca może być całkowicie różna od wypowiedzi testowej (w trakcie próby dostępu do systemu). Ze względu na te same „dane" w trakcie uczenia i testowania poprawność rozpoznawania dla systemów zależnych od tekstu jest większa niż dla niezależnych od tekstu. Ponieważ człowiek rozpoznaje głos niezależnie od zawartości wypowiedzi, znacznie ciekawszym obiektem do badań i analizy są systemy niezależne od tekstu. Ich dodatkową zaletą jest możliwość zabezpieczenia się przed próbami niepowołanego dostępu poprzez

odtworzenie uprzednio nagranego głosu mówcy. Systemy takie bazują na tekście podpowiadanym (*text-prompted*), który mówca musi przeczytać, aby uzyskać ewentualną akceptację. Ze względu na nieograniczoną ilość kombinacji podpowiadanych słów pozwala to na bardzo skuteczne zabezpieczenie się przed próbami odtworzenia wcześniej zarejestrowanego głosu mówcy autentycznego. W tym przypadku jednak system musi dodatkowo rozpoznawać mowę w celu sprawdzenia czy mówca powiedział to o co go poproszono.

W systemie rozpoznawania mówców można wyróżnić kilka głównych elementów składowych. Sygnał akustyczny po spróbkowaniu i podziale na segmenty zwane ramkami jest poddawany operacjom matematycznym, mającym na celu ekstrakcję parametrów sygnałów mowy, które w możliwie największym stopniu przenoszą informację osobniczą o mówcy. Rodzaj tych parametrów i ich ilość mają decydujące znaczenie w procesie rozpoznawania. Zbyt mała ilość ekstrahowanych parametrów powoduje poważne pogorszenie osiągów systemu, przy zbyt dużej z kolei gwałtownie rośnie liczba niezbędnych obliczeń, nie poprawiając w istotny sposób efektywności rozpoznawania. Najbardziej ogólną strukturę systemu identyfikacji mówcy przedstawiono na rys. 1.

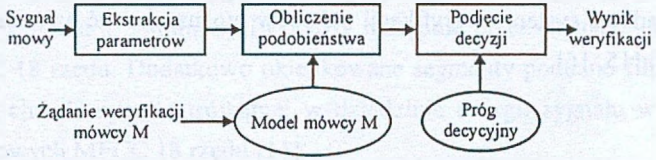


Rys. 1. Ogólna struktura systemu identyfikacji mówcy
Fig. 1. Basic structure of identification system

Parametry ekstrahowane z wypowiedzi osoby rozpoznawanej tworzą ciąg wielowymiarowych wektorów zwanych sekwencją testową. Identyfikacja osoby polega na obliczeniu podobieństwa pomiędzy sekwencją testową a wszystkimi modelami mówców. Osoba zostaje rozpoznana jako ta, dla której sumaryczna odległość pomiędzy jej modelem a ciągiem testowym jest najmniejsza. W przypadku modeli statystycznych zostaje wybrany ten

model, dla którego prawdopodobieństwo wygenerowania zarejestrowanej wypowiedzi jest największe. Taki sposób identyfikacji jest właściwy tylko w przypadku identyfikacji w zbiorze zamkniętym. W przypadku identyfikacji w zbiorze otwartym wprowadza się pewien próg decyzyjny, którego przekroczenie pozwala ustalić, czy osoba rozpoznawana jest już zarejestrowana w systemie czy też nie.

W przypadku systemu weryfikacji mówcy (rys. 2) obliczenie podobieństwa modelu do ciągu testowego wykonuje się tylko dla modelu mówcy, którego tożsamość jest deklarowana.



Rys. 2. Ogólna struktura systemu weryfikacji mówcy
Fig. 2. Basic structure of verification system

W odróżnieniu jednak od identyfikacji w zbiorze zamkniętym istnieje konieczność określenia dla każdego z mówców wartości progu, po przekroczeniu którego zostaje podjęta decyzja o akceptacji bądź odrzuceniu użytkownika. Dodatkową cechą odróżniającą identyfikację od weryfikacji jest wpływ populacji mówców na efektywność działania systemu rozpoznającego. Dla identyfikacji ze wzrostem liczby zarejestrowanych mówców, czyli klas, rośnie monotonicznie prawdopodobieństwo błędnej klasyfikacji, podczas gdy dla weryfikacji jest ono praktycznie stałe [2,3].

3. Zastosowania weryfikacji mówcy

Pomimo pewnych braków, których usunięcie wymaga dalszych badań, technologia weryfikacji mówcy może już obecnie znaleźć wiele zastosowań. Spośród wielu różnych przykładów należy wymienić jej zastosowanie w aplikacjach jako dodatkowa metoda zabezpieczenia przed dostępem nieuprawnionych użytkowników do systemu. Innym zastosowaniem jest bankowość elektroniczna, której rozwój wymaga odpowiednio wysokiego poziomu bezpieczeństwa. Weryfikacja głosu może być tu z powodzeniem zastosowana jako dodatkowe zabezpieczenie. W porównaniu z innymi metodami biometrycznymi ma ona szereg zalet, dzięki którym jest szczególnie dobrze przystosowana do sieci Internet. Najważniejszą z nich jest prostota, gdyż w przeciwieństwie do innych metod identyfikacji, jak odciski palców czy rozpoznawanie twarzy, wystarczy, aby komputer był wyposażony w mikrofon. Szersze wykorzystanie metod biometrycznych w Internecie wymaga jednak opracowania protokołu do transmisji danych biometrycznych, który do tej pory nie istnieje,

jednakże w przypadku głosu problem ten jest bliski rozwiązania [17]. Pokrewnym zastosowaniem do wymienionego jest dostęp do serwisów usługowych poprzez sieć telefoniczną. Weryfikacja głosowa jest do tego celu idealnie przystosowana. Kolejnym zastosowaniem jest kryminalistyka, gdzie często wykonuje się badania fonoskopijne materiałów dowodowych.

Należy jednak podkreślić, że największą zaletą weryfikacji głosowej jest jej stosunkowa prostota, niska cena oraz przystosowanie do obecnych mediów transmisyjnych (telefon, Internet). Przykładowe systemy weryfikacji mówców można znaleźć na stronach www kilku firm i organizacji [15, 16].

4. Implementacja weryfikacji mówcy w środowisku Matlab

W celu sprawdzenia algorytmów weryfikacji mówcy i możliwości jej ewentualnego zastosowania jako dodatkowego zabezpieczenia przed niepowołanym dostępem do systemu komputerowego zaimplementowano w środowisku Matlab procedury identyfikacji i weryfikacji mówcy. Testowanie systemu rozpoznawania mówców powinno być przeprowadzane na standaryzowanych bazach (*speech corpus*), zawierających wypowiedzi wielu mówców nagranych podczas wielu sesji. Instytucjami, które zajmują się tworzeniem takich zasobów mowy, są LDC [7], OGI [10] oraz ELRA [4]. Ponieważ koszt takich zasobów mowy jest dosyć duży, zdecydowano się na wykorzystanie zasobu polskiego ROBOT stworzonego w WAT w Warszawie.

Zasób ROBOT stanowią wypowiedzi 30 mówców obojga płci nagrane w kilku sesjach w celu uchwycenia zmienności głosu ludzkiego na przestrzeni czasu. Cechą charakterystyczną tej bazy są stosunkowo wysoka jakość nagranych wypowiedzi oraz zasób słownictwa ograniczony do liczb oraz komend stosowanych do sterowania robotem (lewo, prawo, złap, puść itd.). Wypowiedzi podzielone są na siedem zbiorów tworzonych w oparciu o pewne reguły. Do stworzenia modeli mówców wykorzystano zbiór Z3 a do celów testowania Z4. Szczegółowy opis tego zasobu można znaleźć w pracy [1].

Skonstruowany system identyfikacji i weryfikacji mówcy umożliwia ekstrakcję kilkunastu różnych parametrów z sygnału mowy (LPC, LPCC, MFCC, delta LPCC, delta MFCC, ACW, LSP, k) [2,3,5,11,18]. Trening systemu (stworzenie modeli mówców) możliwy jest w oparciu o algorytmy LBG i k-średnich [11]. Ponieważ Matlab jest interpreterem, w celu zapewnienia szybkości działania aplikacji algorytmy musiały zostać napisane w postaci zwektoryzowanej, a część z funkcji realizujących najbardziej czasochłonne operacje została skompilowana do postaci bibliotek „dll". Badania przeprowadzono na komputerze z procesorem Celeron 400 MHx i 128 MB RAM.

W trakcie treningu i testowania systemu zastosowano tę samą procedurę przetwarzania sygnału mowy. Pliki dźwiękowe w formacie „wav" wczytywano kolejno do pamięci komputera, po czym w oparciu o kryterium energetyczne usuwano z nich fragmenty ciszy. Następnie sygnał był poddawany preemfazie [11] z parametrem α=0.95 i dzielony na ramki o czasie trwania 10 ms z 50% nakładkowaniem, co dawało 200 ramek na sekundę. Tak uzyskane segmenty sygnału były poddawane okienkowaniu Hamminga, a następnie analizie LPC 12 rzędu, uzyskując tym samym dla każdej ramki 12 współczynników LPC i 12 współczynników odbicia k. Parametry LPC były następnie przekształcane na współczynniki cepstralne LPCC 18 rzędu. Dodatkowo okienkowane segmenty poddano filtracji za pomocą banku filtrów o charakterystyce trójkątnej w dziedzinie energii sygnału w celu uzyskania parametrów melowych MFCC 18 rzędu [11].

Mówca w systemie był reprezentowany poprzez zbiór 2, 4, 8, 16 i 32 wielowymiarowych wektorów kodowych znalezionych w oparciu o algorytm k-średnich. Do stworzenia modeli poszczególnych mówców wykorzystano ich wszystkie wypowiedzi ze zbioru Z3 zasobu ROBOT (75 plików „wav", około 90 s sygnału mowy po usunięciu ciszy). Do testowania wykorzystano wszystkie wypowiedzi ze zbioru Z4, składające się z kombinacji liczb ze zbioru Z3, realizując tym samym rozpoznawanie zależne od tekstu. Ponieważ każdy z mówców dostarczył 11 wypowiedzi, działanie systemu identyfikacji zbadano dla 330 wypowiedzi a skuteczność weryfikacji dla 30*11*30=9900 wypowiedzi. Czas trwania każdej z sekwencji testowych wynosił około 5 s. Zbadano wpływ wielkości modelu i zastosowanych parametrów na skuteczność weryfikacji.

Miernikiem jakości systemu weryfikacji mówcy jest stopa fałszywej akceptacji FA (*False Acceptance Rate*) oraz stopa fałszywego odrzucenia FR (*False Rejection Rate*). Zwykle parametry te służą do wykreślenia krzywych DET [8], które w pełni opisują zachowanie się systemu weryfikacji dla różnych wartości progu decyzyjnego. Innym parametrem chętnie stosowanym do opisu zachowania się systemu jest EER (*Equal Error Rate*), który można znaleźć dobierając wartość progu tak, aby FA było równe FR.

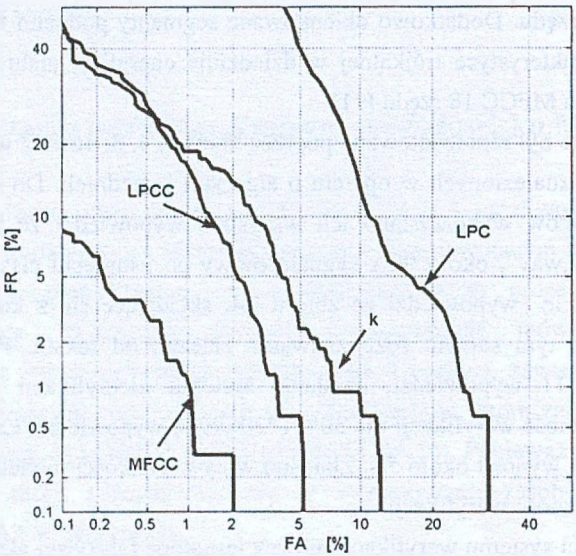
Tabela 1

Stopa EER w % dla różnych parametrów w funkcji wielkości modelu mówcy

Rozmiar modelu	k	LPC	LPCC	MFCC
2	21.49	26.82	13.54	20.18
4	12.93	19.15	8.18	9.84
8	8.93	15.57	5.96	3.88
16	5.96	12.78	3.88	2.24
32	4.14	10.47	2.77	0.98

Wartości EER dla opisywanego systemu weryfikacji zawarto w tab. 1. Można zauważyć, że wraz ze wzrostem liczby wektorów kodowych przypadających na model mówcy maleje wartość tego parametru. Bardziej skomplikowany model wymaga jednak dużo większej

liczby obliczeń. Zdecydowanie najniższy poziom błędów osiągnięto dla parametrów cepstralnych MFCC oraz LPCC. Najlepszy osiągnięty wynik rzędu 1% mówi, że średnio raz na sto prób weryfikacji uprawniony użytkownik zostanie błędnie odrzucony bądź też oszust zostanie błędnie zaakceptowany. Dobierając odpowiednio wartość progu decyzyjnego można po zaakceptowaniu wyższego współczynnika FR uzyskać dużo większą odporność na oszustów. Zachowanie opisywanego systemu weryfikacji mówców w pełni opisuje krzywa DET przedstawiona na rys. 3.



Rys. 3. Osiągi systemu dla różnych parametrów (32 wektory kodowe na mówcę)
Fig. 3. System performance for different parameters (32 code vectors per speaker)

5. Podsumowanie

Weryfikacja mówcy wymaga dalszych badań w celu zmniejszenia wskaźnika błędu EER. W pracy przedstawiono osiągnięte wyniki dla różnych parametrów ekstrahowanych z sygnału mowy jak i dla różnych wielkości modeli mówców. Osiągnięte rezultaty wskazują, że już obecnie technologia weryfikacji głosu ludzkiego może być zastosowana jako metoda autoryzacji przy próbach dostępu do systemów komputerowych, w trakcie logowania się do systemu i innych aplikacjach programowych, gdzie istnieje konieczność zwiększenia bezpieczeństwa pracy systemu informatycznego. Istnieje co prawda pewne prawdopodobieństwo, że uprawniony użytkownik posiadający pełnię praw zostanie podczas

weryfikacji odrzucony, jednakże jest to cecha wszystkich systemów biometrycznych. Weryfikacja głosu ludzkiego w porównaniu z innymi metodami biometrycznymi ma szereg bardzo istotnych zalet, z których można chociażby wymienić prostotę, niską cenę jak i możliwość weryfikacji zdalnej poprzez kanał telefoniczny. Szczególnie ta ostatnia cecha jest bardzo cenna, gdyż prawdopodobnie w niedalekiej przyszłości zwiększy bezpieczeństwo transakcji finansowych dokonywanych w systemach bankowości elektronicznej.

LITERATURA

1. Adamczyk B., Adamczyk K., Trawiński K.: Zasób mowy ROBOT. Biuletyn Instytutu Automatyki i Robotyki WAT, 2000, nr. 12, ss. 179-192.
2. Campbell J. P.: Speaker Recognition: A Tutorial. Proc. IEEE, 1997, vol. 85, no. 9, pp. 1437-1462.
3. Dustor A., Izydorczyk J.: Rozpoznawanie mówców. Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne, 2003, nr 2-3, ss. 71 - 76.
4. European Language Resources Association: http://www.icp.grenet.fr/ELRA/.
5. Furui S.: Cepstral Analysis Techniques for Automatic Speaker Verification. IEEE Trans. Acoustics, Speech, Signal Processing, 1981, vol. 29, pp. 254-272.
6. IBM ViaVoice: http://www-3.ibm.com/software/speech/
7. LDC: http://www.ldc.upenn.edu/
8. Martin A., Doddington G., Kamm T., Ordowski M., Przybocki M.: The DET curve in assessment of detection task performance. Proc. Eurospeech 97, 1997, pp. 1895-1898.
9. Microsoft Speech Technologies: http://www.microsoft.com/speech/download/sdk51/
10. Oregon Graduate Institute: http://cslu.cse.ogi.edu/
11. Rabiner L. R., Juang B. H.: Fundamentals of Speech Recognition. Prentice Hall, 1993.
12. Scansoft: http://www.scansoft.com/
13. Speech Synthesis and Speech Recognition using SAPI5.1: http://www.blong.com/Conferences/DCon2002/Speech/SAPI51/SAPI51.htm
14. Synteza mowy: http://www.syntezatorek.republika.pl/
15. The Biometric Consortium: http://www.biometrics.org/
16. The Speech Technology Center: http://www.speechpro.com/
17. VoiceXML: http://www.voicexml.org/
18. Zilovic M. S., Ramachandran R. P., Mammone R. J.: A Fast Algorithm for Finding the Adaptive Component Weighted Cepstrum for Speaker Recognition. IEEE Trans. Speech Audio Processing, 1997, vol. 5, no. 1, pp. 84-86.

Recenzent: Prof. dr hab. inż. Bolesław Pochopień

Wpłynęło do Redakcji 24 kwietnia 2003 r.

Abstract

The first part of the paper includes fundamentals of speaker identification and verification including differences between these tasks and very brief description of their main parts like feature extraction and model training. Some problems of modern speech technologies are also considered. Applications of speaker verification are discussed in the next part with special emphasis put on access security applications. The rest of the paper describes speaker identification and verification system written in Matlab language. This system can work both in a text-dependent and text-independent mode. However, only text-dependent mode is discussed in this paper. Obtained results are included as well as the influence of parameters extracted from a speech wave and the model size on verification accuracy (Tab. 1). Speaker models were obtained using very well known in speech coding k-means procedure. All research was done utilising Polish speech corpus ROBOT, which contains utterances of 30 speakers collected in a several time-separated sessions to catch intraspeaker variability. The trade off between false rejection rate and false acceptance rate on a DET curve was shown (Fig. 3).

Adres

Adam DUSTOR: Politechnika Śląska, Instytut Elektroniki, ul. Akademicka 16,
44-100 Gliwice, Polska, dustor@iele.polsl.gliwice.pl .