Janusz DULAS
POLITECHNIKA OPOLSKA, INSTYTUT ELEKTROTECHNIKI, AUTOMATYKI I INFORMATYKI
ul. Sosnkowskiego 31, 45-272 Opole

Parametry identyfikacyjne umożliwiające automatyczne rozpoznawanie cyfr wypowiadanych w języku polskim

Dr inż. Janusz DULAS

Dr inż. Janusz Dulas, ur. 4.02.1971r w Opolu. W 1995r ukończył studia na Politechnice Opolskiej, na kierunku Elektrotechnika. W 2002r. obronił rozprawę doktorską pt. „Metoda siatek o zmiennych parametrach w zastosowaniu do rozpoznawania fonemów mowy polskiej". Obecnie adiunkt na Politechnice Opolskiej, na wydziale Elektrotechniki, Automatyki i Informatyki. Zajmuje się badaniami nad automatycznym rozpoznawaniem i sterowaniem za pomocą sygnałów mowy.

e-mail: dulas@po.opole.pl

Streszczenie

Artykuł przedstawia najnowsze wyniki prac autora w dziedzinie automatycznego rozpoznawania sygnałów mowy. Wyniki badań prowadzonych na zbiorze 500 nagrań cyfr wypowiadanych w języku polskim przez 50 mówców różnej płci i w różnym wieku pozwalają na zaproponowanie zestawu parametrów niezbędnych do przeprowadzenia procesu ich identyfikacji. Jak pokazano w artykule zestaw kilku podstawowych cech identyfikujących jest wystarczający aby taki proces przeprowadzić. Zaproponowany zestaw parametrów jest łatwy do uzyskania przy niewielkiej mocy obliczeniowej.

Słowa kluczowe: automatyczne rozpoznawanie sygnału mowy, fonemy.

Identification parameters enabling automatic recognition of digits spoken in Polish

Abstract

The paper describes a new author's method for automatic recognition of digits spoken in Polish. In this new approach there are no frequency analyses as used to be made in such systems but the image recognition of the time characteristic is applied. Investigations performed on 500 records of people of different sex and age showed that there was possibility of constructing an automatic recognition system based on a few parameters. The first is the number of voiced phonemes included in a recognized word (Tab. 1). In this group there are all wavelets and some consonants. They include basic periods inside their time characteristics. This parameter is obtained using the grid method designed by the author (Fig. 3). The second one is the number and position of noisy phonemes. To this group there belong phonemes without basic periods but with big signal variety. This parameter is calculated using the number of local extrema, the signal amplitude level and checking if there are no basic periods. The third parameter is the shape of a signal envelope (Tab. 2). As investigations showed, it is possible to find the envelope pattern for each Polish digit common for all tested speakers. It was proved that these parameters are sufficient for automatic speech recognition of digits spoken in Polish. This new method can also be applied to other systems with small number of recognized words. It is fast and lack of frequency analyses causes that it has low hardware demands.

Keywords: automatic speech recognition, phonemes.

1. Wstęp

Automatyczna identyfikacja sygnałów mowy oraz prace z tym związane są prowadzone od wielu lat [1, 2, 3]. Szybki rozwój techniki komputerowej przyspieszył te prace i obecnie możemy korzystać z coraz większej liczby urządzeń sterowanymi za pomocą sygnałów mowy [4, 5, 6]. Mimo ogromnego postępu prac w tej dziedzinie wciąż jest jednak sporo do zrobienia.

Po pierwsze, kilkaset różnych języków jakimi posługują się ludzie na całym świecie dość skutecznie utrudnia przenoszenie rozwiązań opracowanych dla jednego języka do innych.

Po drugie, w procesie automatycznego rozpoznawania sygnałów mowy mamy do czynienia z żywymi organizmami, z których każdy posiada nieco inną budowę, schorzenia, styl mówienia, nawyki.

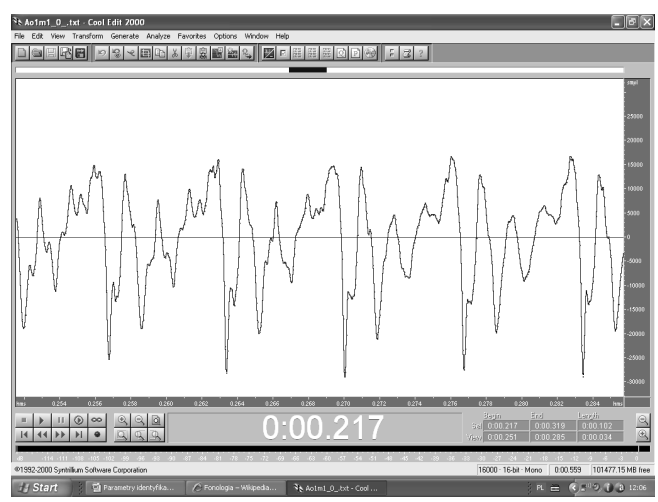
Trzeci problem wynika z faktu, iż ludzie porozumiewając się ze sobą używają nie tylko słów ale również gestów (stąd powstające systemy audiowizualne [7, 8]) oraz „bazy wiedzy", czyli dotychczasowego doświadczenia życiowego umożliwiającego odtworzenie informacji silnie zakłóconej lub nie pełnej.

Kolejnym problemem jest wpływ zakłóceń na pracę systemów ARM (Automatycznego Rozpoznawania Mowy). Stąd liczne prace badawcze dla systemów pracujących w pojazdach [9], czy też identyfikacji mowy w dużych pomieszczeniach z widownią [10, 11].

Szybki rozwój nauki pozwala też na stosowanie różnych metod identyfikacji sygnałów mowy. Do najpopularniejszych należą metody oparte na analizie widmowej, wykorzystujące Łańcuchy Markowa [12, 13, 14, 15] ale spotyka się też rozwiązania wykorzystujące sieci neuronowe, analizy czasowe, metody falkowe i inne. Niniejszy artykuł przedstawia parametry uzyskane z analiz czasowych nagrań mówców różnej płci i różnego wieku wypowiadających izolowane cyfry w języku polskim.

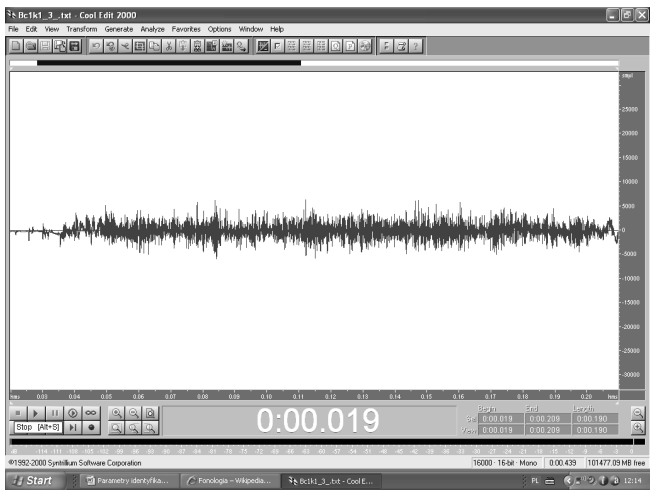
2. Uproszczona analiza fonologiczna cyfr występujących w języku polskim

Wszystkie wyrazy (a więc również i wypowiadane cyfry) zbudowane są z najmniejszych segmentów fonetycznych zwanych fonemami. Dla języka polskiego przyjmuje się zestaw 37 fonemów, który umożliwia zbudowanie 95% wszystkich wyrazów w tym języku [1]. Fonemy można dzielić ze względu na różne kryteria, jednak dla celów dalszej analizy ograniczmy się do rozróżnienia fonemów posiadających okresy podstawowe traktu głosowego (głosowych) oraz fonemów szumowych. Do pierwszej grupy należą wszystkie samogłoski oraz spółgłoski dźwięczne (np. a, e, i, g). Druga grupa to fonemy bezdźwięczne (np. sz, cz, ć). Fonemy pierwszej grupy łatwo jest rozróżnić po występujących w ich przebiegach czasowych okresach periodycznych (rys. 1).



Rys. 1. Przebiegi okresowe w fonemie „o"
Fig. 1. Periodical signal in „o" phoneme

Czasy trwania tych okresów są ściśle związane z budową tonu krtaniowego i są krótsze dla kobiet i dzieci (ok. 2..5 ms) i dłuższe dla mężczyzn (7..10 ms). Jest to ważna cecha osobnicza i umożliwia łatwą identyfikację płci mówcy. Fonemy szumowe nie posiadają okresów podstawowych tonu krtaniowego, cechują się za to dużą zmiennością sygnału i zwykle mniejszą amplitudą (w porównaniu z fonemami pierwszej grupy). Przykładową charakterystykę fonemu szumowego „sz" przedstawia rys. 2.



Rys. 2. Przebieg czasowy fonemu „sz"
Fig. 2. Phoneme "sz" time characteristic

Stosując podział fonemów na w/w grupy można zapisać strukturę fonemową każdej z cyfr stosując oznaczenie "1" dla fonemu z okresem podstawowym i "S" dla fonemu szumowego, co przedstawia tabela 1.

Tab. 1. Struktura fonemowa cyfr w języku polskim
Tab. 1. The phoneme structure for digits in Polish

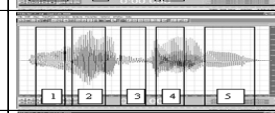
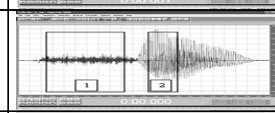
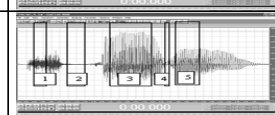
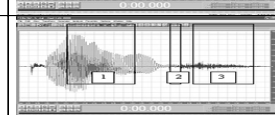
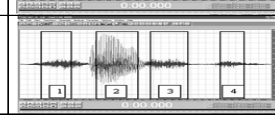
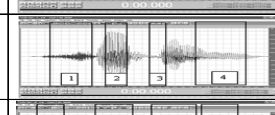
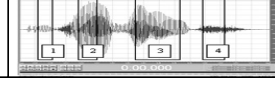
Cyfra	Cyfra słownie	Struktura fonemowa
0	ZERO	1,1,1
1	JEDEN	1,1,1,1
2	DWA	1,1,1
3	TRZY	S,1
4	CZTERY	S,1,1
5	PIĘĆ	1,1,S
6	SZEŚĆ	S,1,S,S
7	SIEDEM	S,1,1,1
8	OSIEM	1,S,1,1
9	DZIEWIĘĆ	1,1,1,1,1,S

3. Analiza obwiedni

Jak wynika z tabeli 1 cyfry „zero" oraz „dwa" mają taką samą strukturę, konieczne jest więc wprowadzenie jeszcze jednego parametru identyfikacyjnego. Może nim być dopasowanie do wzorca obwiedni. W badaniach przeprowadzonych przez autora [15, 16] udało się stworzyć zestaw 10 wzorców obwiedni (jeden dla każdej z cyfr), które były wspólne dla wszystkich badanych mówców (badania przeprowadzono dla 500 nagrań). Każdy z wzorców podzielono na fragmenty oraz opisano za pomocą zakresu amplitud sygnału i minimalnego czasu trwania Zestawienie wszystkich wzorców przedstawia tabela 2.

Jak wynika z poniższej tabeli, cyfry „0" i „2" mają różną liczbę fragmentów, co może być wykorzystane przy ich identyfikacji. Dokładny opis parametrów każdego z wzorców został zamieszczony w [16].

Tab. 2. Wzorce obwiedni dla cyfr od 0 to 9
Tab. 2. Envelope patterns for digits from 0 to 9

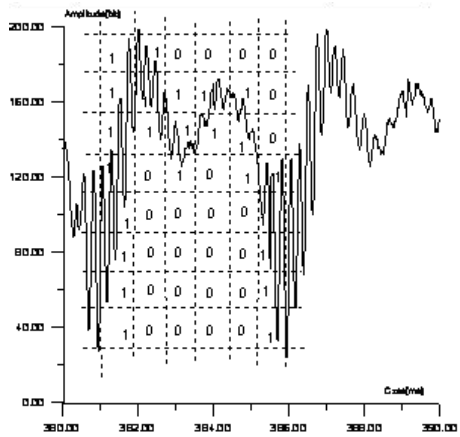
Cyfra	Kszat obwiedni z zaznaczona liczbą charakterystycznych fragmentów	Liczba fragmentów	Minimalne czasy trwania	Zakresy amplitud
0		4	1.80ms 2.30ms 3.10ms 4.50ms	1.A<50% 2.A>53% 3.5-64% 4.A>31%
1		5	1.60ms 2.70ms 3-50ms 4-60ms 5-100ms	1.A<62% 2.A>44% 3-1%<A<26% 4-A>25% 5-A<50%
2		3	1.50ms 2.20ms 3.50ms	1.A<30% 2.A>29% 3.A>43%
3		2	1.130ms 2.40ms	1.6%<A<32% 2.A>68%
4		5	1.30ms 2.40ms 3.100ms 4.10ms 5.60ms	1.A>1% 2.A<4% 3.A>40% 4.3%<A<35% 5.A>13%
5		3	1.140ms 2.20ms 3.110ms	1.A>29% 2.A<23% 3.A>0%
6		4	1.80ms 2.100ms 3.100ms 4.70ms	1.2%<A<35% 2.A>26% 3.1%<A<56% 4.A>1%
7		4	1.100ms 2.50ms 3.30ms 4.110ms	1.3%<A<65% 2.A>36% 3.2%<A<28% 4.6%<A<59%
8		4	1.70ms 2.90ms 3.100ms 4.90ms	1.A>46% 2.6%<A<61% 3.16%<A<92% 4.5%<A<43%
9		4	1.40ms 2.70ms 3.150ms 4.80ms	1.1%<A<45% 2.A>46% 3.A>9% 4.A<50%

4. Wykrywanie fonemów posiadających okresy podstawowe tonu krtaniowego metodą siatek

Jak już wspomniano w punkcie 2 okresy podstawowe tonu krtaniowego są periodycznymi, wielokrotnie powtarzającymi się przebiegami o okresie od 2 do 10ms. Ponieważ czasy trwania fonemów z tej grupy wahają się od kilkudziesięciu do kilkuset milisekund, w tym czasie można odnaleźć od kilkunastu do kilkudziesięciu okresów podstawowych. Część z nich cechuje się dużym podobieństwem co zostało wykorzystane do celów identyfikacyjnych. Aby określić stopień podobieństwa kolejnych okresów nakłada się na każdy z nich siatkę o pewnej, stałej rozdzielczości i w jej komórki wpisuje się „0" (tam, gdzie sygnału nie ma) oraz „1" (gdzie sygnał występuje). Ważne jest tu aby szerokość siatki była dopasowana do czasu trwania jednego okresu, a jej wysokość do amplitudy sygnału. Zasadę tą obrazuje rys. 3.

Porównując te same komórki siatek dla sąsiadujących okresów podstawowych można określić stopień ich podobieństwa rozumiany jako liczba „1" na tych samych pozycjach siatek. W algorytmie identyfikacyjnym przyjęto, iż 88% lub więcej zgodnych jedynek oznacza podobieństwo siatek. Analizując dany fonem można takie porównania przeprowadzić dla każdego okresu podstawowego, porównując uzyskaną dla niego siatkę z fonemami poprzedzającymi i następującymi po nim. W trakcie badań przyjęto, iż dokonuje się porównań dla pięciu siatek poprzedzających i pięciu następujących po badanej siatce. W ten sposób dla każdego okresu obliczono współczynnik podobieństwa, którego wartość może wahać się od 1 (badana siatka nie jest do żadnej podobna, poza

sobą samą) do 11 (badana siatka jest podobna do 5 poprzedzających, 5 następujących po niej i do siebie samej).



Rys. 3. Dopasowanie siatki do okresu podstawowego
Fig. 3. Grid fitting to the basic period

Przykładowe zestawienie współczynników podobieństwa dla wyrazu „trzy” przedstawiono na rysunku 4.

2,1,2,3,2,1,3,1,1,2,2,2,1,1,2,1,1,3,1,3,1,3,5,4,3,6,4,1,2,3,3,3,3,1,3,2,1,2,3,1,5,1,4,1,
1,3,2,2,1,2,1,2,2,2,1,4,2,3,2,1,2,3,2,1,1,1,3,4,4,4,1,1,2,3,3,4,2,3,2,3,1,3,2,3,2,4,3,7,
6,6,7,9,10,9,9,10,11,11,10,11,11,10,10,9,8,10,10,5,8,3,5,5,4,1,2,2,2,3,5,4,3,3,4,3,

Rys. 4. Współczynniki zgodności siatek dla wyrazu „trzy”
Fig. 4. Similarity coefficients for digit 3

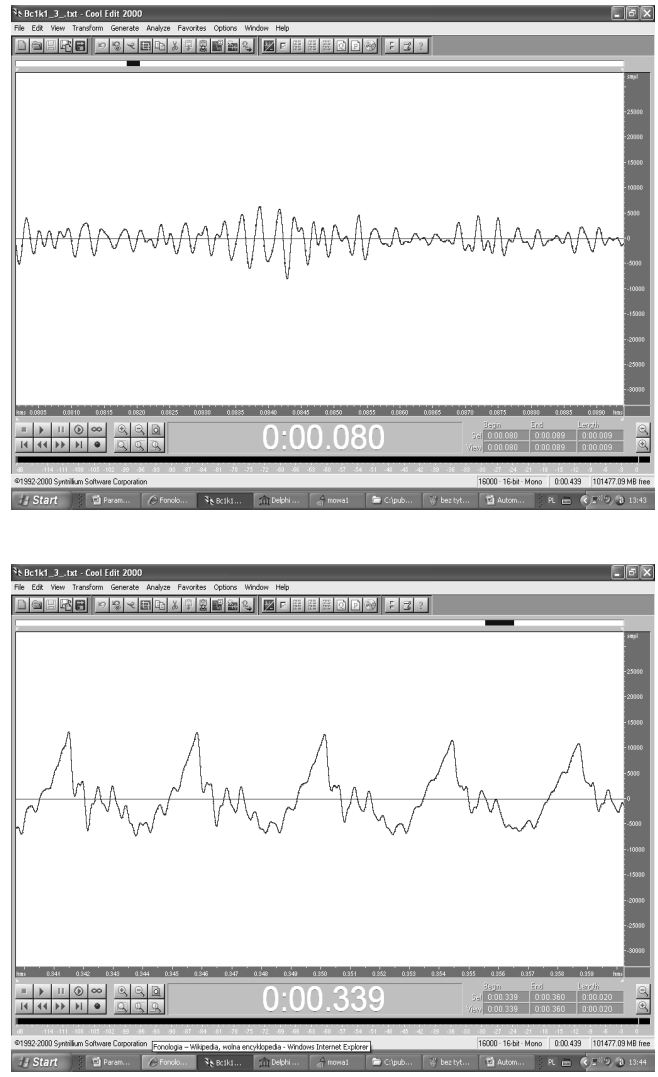
Jak wynika z powyższego rysunku duża liczba siatek o wysokich współczynnikach zgodności skupiona jest w drugiej części wyrazu, co wyraźnie sygnalizuje znajdowanie się tam fonemu posiadającego okresy podstawowe tonu krtaniowego. W tym przypadku jest to fonem „y”. Jest to zgodne z przewidywaną teoretycznie liczbą takich fonemów co przedstawiono w tabeli 1 w wierszu dla cyfry „3”. Z rysunku 3 wynika również, iż w początkowej części wyrazu też znajduje się fonem (fonemy?) lecz o małych współczynnikach zgodności co oznacza iż nie są to samogłoski ani spółgłoski dźwięczne (gdyż brak jest tam okresów podstawowych tonu krtaniowego).

5. Wykrywanie fonemów szumowych

Fonemy szumowe charakteryzują się dużą zmiennością sygnału (patrz rys.2), brakiem okresów podstawowych tonu krtaniowego oraz zwykle mniejszą amplitudą niż fonemy dźwięczne. Te obserwacje pozwoliły na zbudowanie algorytmu wykrywającego fonemy szumowe. Aby wykryć dużą zmienność sygnału wyznaczana jest liczba ekstremów lokalnych w przeliczeniu na czas trwania fonemu. Liczba ta jest zwykle wyższa niż dla fonemów nie szumowych. Przedstawia to rys. 5, gdzie w czasie 10ms fonemu szumowego (po lewej stronie rysunku) zaobserwowano 48 maksimów lokalnych, a w tym samym czasie dla fonemu z okresami podstawowymi (po prawej stronie rysunku) 37 maksimów lokalnych.

Drugą cechą charakterystyczną dla fonemów szumowych jest brak występowania okresów podstawowych tonu krtaniowego. Można to stwierdzić przez analizę współczynników podobieństwa siatek co zostało opisane w poprzednim punkcie.

Kolejną cechą braną pod uwagę jest zakres amplitud dla danego fonemu. Konieczne jest tu zwłaszcza ustalenie dolnej granicy na poziomie kilku procent maksymalnej amplitudy sygnału występującego w danym wyrazie aby wyeliminować pomyłki związane z interpretowaniem sygnału ciszy jako fonemu szumowego.



Rys. 5. Różne liczby ekstremów lokalnych dla fonemów szumowych i nie szumowych
Fig. 5. Different numbers of local extrema for noisy and unnoisy phonemes

6. Przebieg i wyniki badań

Jak wspomniano na wstępie do badań wykorzystano 500 nagrań cyfr z zakresu od 0 do 9 wypowiadanych przez mówców różnej płci i w różnym wieku (50 nagrań dla każdej z cyfr). Początkowo wyznaczono wzorce obwiedni. Na podstawie obserwacji charakterystyk czasowych tej samej cyfry dla różnych mówców podzielono każde nagranie na fragmenty wyznaczając ich minimalny czas trwania i zakres amplitudy sygnału. Po wstępnym określeniu wzorców przeprowadzono ich weryfikację dla wszystkich mówców zmieniając w razie potrzeby oba parametry każdego z fragmentów tak aby były one zgodne dla każdego z nagrań. Stworzono tablicę wzorców, która jest prawdziwa dla wszystkich 500 nagrań.

W celu identyfikacji fonemów posiadających okresy podstawowe tonu krtaniowego opracowano metodę siatek oraz algorytm obliczający współczynniki podobieństwa dla każdej z siatek. Otrzymane współczynniki umożliwiają wykrywanie takich fonemów. Badania tego algorytmu przetestowano dla cyfr z zakresu od 0 do 9.

Bazując na omówionych wcześniej parametrach, stworzono algorytm wykrywający fonemy szumowe. Jego działanie przetestowano na cyfrach z zakresu od 0 do 9.

Otrzymane wyniki badań zamieszczono w tabeli 3.

Tab. 3. Wyniki badań
Tab. 3. Investigation results

Cyfra	Liczba poprawnie rozpoznanych nagrań tylko za pomocą analizy obwiedni	Liczba poprawnie rozpoznanych nagrań za pomocą analizy obwiedni, wykrywania fonemów szumowych i okresów podstawowych tonu krtaniowego
-	[%]	[%]
0	82	100
1	34	100
2	52	100
3	22	100
4	6	100
5	36	100
6	22	100
7	100	100
8	42	100
9	22	100
Średnia:	**41,8**	**100,0**

Jak wynika z powyższej tabeli zaproponowany zestaw parametrów umożliwia pełną identyfikację cyfr dla wszystkich mówców.

7. Podsumowanie

W pracy przedstawiono aktualny stan badań autora w dziedzinie automatycznej identyfikacji sygnałów mowy. Badania przeprowadzono na zestawie 500 nagrań cyfr wypowiadanych w języku polskim przez mówców różnej płci i w różnym wieku. Otrzymany zestaw parametrów obejmujący analizę obwiedni, wykrywanie fonemów szumowych oraz zliczanie fonemów posiadających okresy podstawowe tonu krtaniowego okazał się wystarczający do bezbłędnej identyfikacji cyfr w zakresie od 0 do 9. Wyniki badań pokazują, iż opracowana przez autora metoda wykazuje lepsze wyniki rozpoznawania i krótszy czas analizy niż metody już funkcjonujące [13].

8. Literatura

[1] Basztura Cz.: Rozmawiać z komputerem, wydawnictwo Format, Wrocław 1992.
[2] Łobacz P., Mikołajczak N., Wysocka J.: Psychofonetyczne podstawy segmentacji sygnału mowy, Prace IPPT, Warszawa 1990.
[3] Tadeusiewicz R.: Sygnał mowy, WKiŁ, Warszawa 1987.
[4] http://simblog.pl/programy-do-sterowania-telefonem-za-pomoca-glosu
[5] http://media2.pl/telekomunikacja/47734-microsoft-mobile-sterowany-glosem.html
[6] http://www.hub30.com/artykul/1694,1,Profesjonalne-dyktafony-dla-pracy-i-rozrywki-ze-sterowaniem-glosowym-i-trybem-PCM
[7] Seymour R., Steward D., Ming J.: Audio-visual integration for robust speech recognition using maximum weighted stream posteriors, INTERSPEECH 2007, Antwerpia, Belgia, 654-657.
[8] Bekiarski A., Pleshkova-Bekiarska S.: Pomiar sygnału głosowego za pomocą matrycy mikrofonowej dwuwymiarowej przeznaczonej do audio-wizyjnego sterowania robota, PAK 10/2008, 741-743.
[9] Weifeng L., Herve B., Non-linear spectra kontrast stretching for In-car speech recognition, INTERSPEECH 2007, Antwerpia, Belgia, 1122-1125.
[10] Lamel L., Adda G., Bilinski E., Gauvain J.L.: Transcribing lectures and seminars, INTERSPEECH 2005, Lisbon, Portugal, 1657-1660.
[11] Trancoso I., Nunes R., Neves L.: Recognition of classroom lectures in european Portuguese INTERSPEECH 2006, Pittsburgh, USA, 281-284.
[12] Juho P., Hanseok K.: A New state-dependent phonetic tied-mixture model with head-body-tail structured HMM for Real time continous phoneme recognition system, INTERSPEECH 2006, Pittsburgh, USA, 1583-1586.
[13] Wydra S.: Recognition quality improvement In automatic speech recognition system for Polish, EUROCON 2007,Warszawa, 218-223.
[14] Kant C., Nishimoto T., Sagayama S.: Model adaptation by state splitting of HMM for long reverberation, INTERSPEECH 2005, Lisbona, Portugalia, 277-280.
[15] Aboutabit N., Beautemps D., Clarke J., Besacier L.: A HMM recognition of consonant-vowel syllables from lip contours: the cued speech case, INTERSPEECH 2007, Antwerpia, Belgia, 646-649.
[16] Dulas J.: Automatyczna segmentacja sygnałów mowy w oparciu o metodę siatek o zmiennych parametrach, PE 1/2010, 229-232.
[17] Dulas J.: Analiza obwiedni jako parametr wspomagający automatyczną identyfikację wyrażeń, PAK 5/2009, 308-309.

otrzymano / received: 25.11.2010
przyjęto do druku / accepted: 02.02.2011 *artykuł recenzowany*

INFORMACJE

Newsletter PAK

Wydawnictwo PAK wysyła drogą e-mailową do osób zainteresowanych Newsletter PAK, w którym są zamieszczane:
- spis treści aktualnego numeru miesięcznika PAK,
- kalendarz imprez branżowych,
- ważniejsze informacje o działalności Wydawnictwa PAK.

Newsletter jest wysyłany co miesiąc do osób, które w jakikolwiek sposób współpracują z Wydawnictwem PAK (autorzy prac opublikowanych w miesięczniku PAK, recenzenci, członkowie Rady Programowej, osoby które zgłosiły chęć otrzymywania Newslettera).

Celem inicjatywy jest umocnienie w środowisku pozycji miesięcznika PAK jako ważnego i aktualnego źródła informacji naukowo-technicznej.

Do newslettera można zapisać się za pośrednictwem:
- strony internetowej: www.pak.info.pl, po dodaniu swojego adresu mailowego do subskrypcji,
- adresu mailowego: wydawnictwo@pak.info.pl, wysyłając swoje zgłoszenie.

Otrzymywanie Newslettera nie powoduje żadnych zobowiązań ze strony adresatów. W każdej chwili można zrezygnować z otrzymywania Newslettera.

Tadeusz SKUBIS
Redaktor naczelny Wydawnictwa PAK