AUDIOFONOLOGIA 2005, 27

Problemy teorii i praktyki
Problems of theory and practice

System rozpoznawania mowy wykorzystujący cechy wizualne

Speech recognition system based on visual features

Piotr Dalka[1], Bożena Kostek[1,2]

[1] Politechnika Gdańska
[2] Centrum Doskonałości „Słuch i Mowa – PROKSIM", Warszawa

Streszczenie

Niniejszy artykuł przedstawia metodę rozpoznawania głosek na podstawie analizy ruchu ust, która może znaleźć zastosowanie w terapii logopedycznej osób z wadami słuchu. W pracy opisano algorytm wyznaczana i śledzenia położenia ust oraz zbadano efektywność jego działania. Sztuczna sieć neuronowa została wykorzystana jako klasyfikator rozpoznający sześć wypowiadanych samogłosek w oparciu o wizualne parametry mowy. Dodatkowo przeprowadzono badania dotyczące rozpoznawania mowy w oparciu o parametry wizualne i akustyczne łącznie. W pracy umieszczono wyniki eksperymentów oraz pokrótce przedstawiono założenia aplikacji komputerowej wspomagającej osoby niedosłyszące.

Słowa kluczowe: odczytywanie mowy z ust, automatyczne rozpoznawanie mowy, sieci neuronowe.

Summary

This article describes a speech recognition system based on the lip movement analysis, designed for hearing impaired. The algorithm for locating and tracing lip movements in video recordings is presented and evaluated. Visual speech features are extracted and fed to the Artificial Neural Networks the task of which is to classify six Polish vowels. Additional experiments were carried out based both on visual and acoustical features. The results of the performed experiments and conclusions are included in the paper. The application of such a system to hearing impaired people is also outlined.

Key words: lipreading, automatized speech recognition, neural networks.

Wprowadzenie

Automatyczne rozpoznawanie mowy (ARW) zyskuje coraz więcej zastosowań. Istniejące obecnie systemy rozpoznawania mowy bazują z reguły na informacji akustycznej i z tego względu są bardzo wrażliwe na szumy otoczenia oraz często zawodzą w przypadku, gdy wiele osób mówi jednocześnie. Błędnie sklasyfikowane sygnały mogą być do pewnego stopnia skorygowane przy wykorzystaniu wysokopoziomowej informacji kontekstowej wykorzystującej reguły gramatyczne i bazy słownikowe, jednakże takie działanie nie jest możliwe w przypadku rozpoznawania nazwisk, adresów i innych nazw własnych. Z kolei ludzie nie mają problemu ze zrozumieniem takich słów, gdyż podświadomie wykorzystują dodatkowe informacje pochodzące bezpośrednio od mówcy, takie jak ruchy ust, pozycja języka i widoczność zębów. W rzeczywistości percepcja mowy ma charakter dwumodalny i informacja wizualna jest często równie istotna, jak dźwiękowa [Dodd, Campbell 1987]. Świadczy o tym fakt, że część osób niesłyszących potrafi czytać z ruchu ust oraz to, że brak kontaktu wzrokowego z mówcą może utrudnić zrozumienie jego wypowiedzi. Większość informacji wizualnej związanej z mową zawarta jest przede wszystkim w ruchach ust oraz w mniejszym stopniu w widoczności języka i zębów [Summerfield 1992]. Z tego względu naturalne wydaje się wykorzystanie analizy obrazu w systemach rozpoznawania mowy.

Celem przeprowadzonych eksperymentów jest stworzenie systemu wyznaczania i analizy wizualnych cech mowy, który może znaleźć zastosowanie w audiologii i terapii logopedycznej osób z uszkodzonym słuchem. System taki mógłby wspomagać logopedów w ćwiczeniach z osobami niedosłyszącymi, bądź osobami po operacji wszczepienia implantów ślimakowych. Zaletą takiej aplikacji komputerowej może być zapewnienie materiału ćwiczeniowego dostosowanego do konkretnych zaburzeń, ułatwienie ćwiczeń, przyspieszenie korekcji zaburzeń mowy, odciążenie głosowe logopedy, itp. poprzez funkcję rozpoznawania mowy, która pozwala ćwiczyć rozumienie i wymowę w oparciu o wzór poprawnej głoski, fonemu lub innej jednostki leksykalnej. Założenia takiego systemu zostały przedstawione przez autorów niniejszego artykułu na Konwencji Audio Engineering Society w Barcelonie [Kostek, Dalka 2005].

Wszystkie eksperymenty opisane w niniejszym referacie zostały przeprowadzone w oparciu o bazę przygotowanych uprzednio 108 nagrań (każde o długości od 1 do 2 sekund) dziewięciu mówców, z których każdy dwukrotnie wypowiadał sześć polskich samogłosek (*a*, *e*, *i*, *o*, *u*, *y*). Nagrania zostały zapisane w standardzie DV (obraz w pełnej rozdzielczości

PAL przy 25 ramkach na sekundę, dźwięk z częstotliwością próbkowania 48 kHz i rozdzielczością 16 bitów). Mówcy w bazie różnili się ze względu na kształt ust, makijaż i zarost. Opisana aplikacja stanowi zaczątek systemu, który w przyszłości będzie wykorzystywał pełną bazę nagrań dźwięków języka polskiego.

Modele ust

W niniejszej pracy proces wyznaczania wizualnych parametrów mowy z nagranych sekwencji filmowych oparty jest na dwóch modelach ust, będących implementacją metody *Acitive Shape Models* (*ASM*) [Cootes (i in.) 1995]. Są to elastyczne modele, w których obiekt jest reprezentowany przez zestaw punktów umieszczonych na krawędzi obiektu lub w jego innych charakterystycznych punktach. Informacja o możliwych zmianach kształtu modelowanego obiektu jest nabywana w drodze analizy statystycznej zestawu treningowego. Na obrazach wchodzących w skład tego zestawu położenie i kształt obiektu jest zaznaczany ręcznie. Unika się w ten sposób heurystycznych założeń dotyczących dozwolonych zmian kształtu ust.

Definicja modeli

Opracowano dwa różne modele ust. Pierwszy z nich (*M1*) opisuje krawędzie zewnętrze górnej i dolnej wargi. Drugi z nich (*M2*) opisuje zarówno krawędź zewnętrzną, jak i wewnętrzną obu warg. Modele te są niezależne od liniowych przekształceń geometrycznych obrazu, wobec czego do wygenerowania konkretnego kształtu konieczna jest również informacja o przesunięciu (t_x, t_y), obrocie (Θ) i skali (s). W przypadku modelu *M1* wykorzystano 22 punkty rozłożone równomiernie na zewnętrznych krawędziach ust; dla modelu *M2* wykorzystano 36 punktów – 22 na krawędziach zewnętrznych i 14 na wewnętrznych.

Przygotowanie zestawu treningowego

W skład zestawu treningowego weszło 216 obrazów, po dwa z każdego nagrania (pierwsza i środkowa ramka). W *M*-obiektowym zbiorze treningowym *i*-ty kształt ust, składający się z *N* punktów, może zostać opisany w postaci wektora [Cootes (i in.) 1995]:

$$x_i = [x_{i1} \quad y_{i1} \quad x_{i2} \quad y_{i2} \quad \dots \quad x_{iN} \quad y_{iN}]^T, \quad i = 1 \dots M \tag{2.1}$$

Na model ust składa się informacja o kształcie średnim oraz o typowych zmianach kształtu ust w porównaniu z kształtem średnim, które uzyskuje się z analizy statystycznej kształtów ust w zestawie treningowym. Aby taka analiza miała sens, konieczne jest wcześniejsze zmniejszenie wpływu różnic w położeniu, obrocie i skali ust. Zwykła normalizacja każdego kształtu z zestawu treningowego do jednostkowej szerokości, zerowego obrotu i zerowego przesunięcia może skutkować wprowadzaniem sztucznych zależności zakłócających tworzony model. Zamiast tego zastosowano proces iteracyjny, w którym każdy kształt był dopasowywany do kształtu średniego (w pierwszym kroku: do dowolnie wybranego kształtu ze zbioru treningowego), a następnie normalizowany był jedynie kształt średni [Cootes (i in.) 1995]. Dopasowanie dwóch kształtów do siebie polega na znalezieniu takich wartości przesunięcia t_x, t_i, obrotu Θ i skali s, aby zminimalizować różnicę dwóch wektorów opisujących kształty. Dodatkowo zastosowano wagi, aby nadać większe znaczenie w procesie dopasowywania tym punktom, które są najbardziej stabilne w zestawie treningowym, tzn. takim, które się najmniej poruszają względem pozostałych punktów składających się na kształt.

Po zakończeniu opisanego wyżej procesu otrzymuje się kształt średni oraz zestaw *N* kształtów treningowych przygotowanych do analizy statystycznej.

Analiza statystyczna

Analiza statystyczna dopasowanych kształtów z zestawu treningowego wykorzystuje Analizę Składowych Głównych (ang. *Principal Componant Analysis*, *PCA*) [Jackson 1991] macierzy kowariancji. Macierz tę można wyznaczyć ze wzoru:

$$S_{2Nx2N} = \frac{1}{M} \sum_{i=1}^{M} (x_i - \bar{x})(x_i - \bar{x})^T \tag{2.2}$$

gdzie:
$\bar{x}$ – kształt średni,
M – liczba obiektów.

W wyniku analizy *PCA* macierzy kowariancji *S* uzyskuje się macierz *2Nx2N* zawierającą wektory własne oraz *2N* związanych z nimi wartości własnych. Wektory własne mają jednostkową długość i są ortogonalne. Wektory własne macierzy kowariancji posiadające największe wartości własne odpowiadają najbardziej znaczącym modom zmienności kształtu. Jednocześnie wariancja związana z każdym wektorem własnym jest równa jego wartości własnej.

Znormalizowany kształt może być na tej podstawie aproksymowany za pomocą wyrażenia [Luettin, Thacker 1997]:

$$x = \bar{x} + P \cdot b \tag{2.3}$$

gdzie:
P – macierz $[p_1 \ p_2 \ \dots \ p_t]$, gdzie $t<2N$, pierwszych t wektorów własnych cechujących się największymi wartościami własnymi,
b – wektor zawierającym wagi dla każdego wektora własnego.

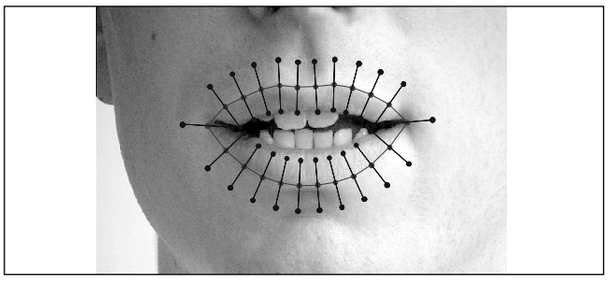
Liczba wektorów własnych użytych do opisu modów kształtu ust jest zwykle znacznie mniejsza od liczby zmiennych opisujących kształt ust.

Lokalizowanie ust

Aby wykorzystać modele kształtu ust do wyznaczenia położenia i kształtu ust w analizowanym obrazie, konieczne jest określenie sposobu pomiaru dopasowania modelu ust do rzeczywistych konturów ust w obrazie. Do tego celu wykorzystano dwa modele luminancji ust (po jednym dla każdego modelu kształtu), bazujące na rozkładzie odcieni szarości wokół konturów ust.

Dla każdego obrazu w zestawie treningowym (tożsamym z zestawem dla wyznaczania modelu kształtu) wyznaczono *N* profili luminancji. Profil g_{ij} jest N_p-elementowym wektorem zawierającym wartości odcieni szarości w punktach rozłożonych równomiernie na odcinku prostopadłym do konturu ust, zaczepionym w *j*-tym punkcie konturu na *i*-tym obrazie (ryc. 1). Zamiast wyznaczania średnich profili luminancji i macierzy kowariancji dla każdego punktu kształtu oddzielnie, stworzono globalny profil luminancji h_i dla *i*-tego obrazu w zbiorze treningowym, będący *N* N_p-elementowym wektorem postaci [Luettin, Thacker 1997]:

$$h_i = [g_{i1} \quad g_{i2} \quad \dots \quad g_{iN}]^T, \quad i = 1 \dots M \tag{3.1}$$



Ryc. 1. Wyznaczanie profilu luminancji dla modelu M1

Następnie analogicznie, jak w przypadku modelowania kształtu, obliczany jest globalny profil średni $\bar{h}$ oraz macierz kowariancji, na podstawie której dokonuje się analizy PCA. Dowolny profil luminancji może być wówczas aproksymowany wyrażeniem:

$$h = \bar{h} + P_g \cdot b_g \tag{3.2}$$

gdzie:
P_g – macierz $[p_{g1}\ \ p_{g2}\ \ \ldots\ \ p_{gt}]$, gdzie $t < N$ N_p, pierwszych t wektorów własnych cechujących się największymi wartościami własnymi,
bg – wektor zawierającym wagi dla każdego wektora własnego.

Proces lokalizacji ust w obrazie jest iteracyjny i polega na minimalizacji funkcji kosztu opisującej dopasowanie modelu luminancji „rozpiętego" na określonym kształcie ust do analizowanego obrazu. Mając określony pewien model kształtu *x* i jego lokalizację na obrazie, wagi bg wektorów własnych profilu luminancji *h* odpowiadającego kształtowi *x*, można wyznaczyć przekształcając wzór (3.2) do poniższej postaci:

$$b_g = P_g^T \left(h - \bar{h}\right) \tag{3.3}$$

Funkcję kosztu definiuje się jako błąd Ep [Luettin, Thacker 1997]:

$$E_p = \left(h - \bar{h}\right)^T \left(h - \bar{h}\right) - b_g^T b_g \tag{3.4}$$

Błąd E_p jest tym mniejszy, im bardziej profil luminancji *h* jest podobny do statystycznego profilu luminancji uzyskanego z analizy zestawu treningowego (wektor $\bar{h}$ i macierz P_g), czyli tym mniejszy, im dokładniej model kształtu ust *x* pokrywa się z rzeczywistym konturem ust w obrazie.

Do znalezienia minimum funkcji błędu E_p wykorzystany został algorytm *Downhill Simplex Method* [Nelder, Mead 1965]. Zmiennymi niezależnymi procesu optymalizacyjnego są przesunięcia t_x i t_y, skala *s*, obrót Θ oraz parametry kształtu *b*. Algorytm ten jest powszechnie stosowany i skuteczny, a przy tym charakteryzuje się przystępną geometryczną interpretacją działania, przez co doskonale nadaje się do prowadzenia badań. Wadą tego algorytmu jest natomiast długi czas potrzebny na otrzymanie wyniku ze względu na wielokrotne częstsze obliczanie wartości minimalizowanej funkcji w porównaniu z innymi algorytmami.

Przyjęto, że obszar obrazu, w którym znajdują się usta jest z góry znany. Model kształtu jest inicjowany kształtem średnim (wektor *b=0*) umieszczonym w wybranym miejscu obszaru poszukiwań. Następnie obliczany jest wektor b_g parametrów luminancji w miejscu, gdzie znajduje się model kształtu (3.3) oraz wyznaczana jest wartość E_p funkcji kosztu (3.4). Algorytm iteracyjnie zmienia kształt ust (wektor *b*) oraz jego położenie na obrazie (zmienne t_x i t_y, *s*, Θ), tak aby zminimalizować wartość E_p. Zakończenie procesu iteracji następuje w chwili, gdy różnica między kolejnymi wartościami E_p spadnie poniżej określonego progu.

Możliwe deformacje kształtu ust zostały ograniczone w ten sposób, że poszczególne mody ust muszą się mieścić w przedziale ±3 odchyleń standardowych, co odpowiada 99% wszystkich możliwych kształtów przy założeniu rozkładu Gaussa.

W przypadku sekwencji obrazów punktem startowym do poszukiwania ust w bieżącej ramce jest lokalizacja ust w ramce poprzedniej.

Eksperymenty

Na podstawie wcześniejszych badań przyjęto rozmiar pojedynczego wektora luminancji N_p równy 21 punktów. W przypadku modelu *M1* wykorzystano 17 pierwszych modów kształtu, a w przypadku modelu *M2* – 28 pierwszych modów, co w obu przypadkach uwzględniało 99,9% możliwych kształtów.

Do opisu modeli luminancji wykorzystano 9 pierwszych wektorów własnych dla modelu *M1* i 14 pierwszych wektorów własnych dla *M2*, co stanowiło 80% możliwych rozkładów luminancji.

Szczególnie krytyczne znaczenie dla skuteczności określania położenia ust ma ilość użytych modów luminancji: zbyt duża lub zbyt mała ich liczba sprawia, że proces minimalizacji funkcji kosztu E_p kończy się przedwcześnie w lokalnym minimum, skutkując błędem lokalizacji.

Wyniki lokalizacji ust

Problemem związanym z oceną skuteczności działania algorytmu lokalizacji ust jest brak obiektywnej, automatycznej metody jej przeprowadzania. Określenie poprawności lokalizacji wymaga przejrzenia sekwencji wynikowych, stąd ocena jest czysto subiektywna i mimo przyjętych ścisłych kryteriów niekiedy trudno o powtarzalność wyników.

W tab. 1 umieszczono wyniki lokalizacji ust w nagraniach z wykorzystaniem obu przygotowanych modeli. Ocenie podlegała nagrana sekwencja jako całość, a nie pojedyncze jej ramki. Oceną 3 oznaczono te nagrania, w których w każdej ramce kształt ust został określony w sposób idealny. Ocenę 2 otrzymały te nagrania, w których w choć jednej ramce wystąpiły pewne rozbieżności znalezionego kształtu w stosunku do rzeczywistego konturu, lecz były one na tyle nieznaczne, że nie powinny mieć wpływu na proces rozpoznawania mowy. Niższą oceną (1) zostały oznaczone te nagrania, w których w przynajmniej jednej ramce występują znaczne błędy w odwzorowaniu konturu ust. Na najniższą ocenę zasłużyły te nagrania, w których pojawiają się poważne, dyskwalifikujące błędy związane z nieprawidłowym określeniem położenia ust, np. zlokalizowanie ich na brodzie mówcy. Oceny 2 i 3 uznano za satysfakcjonujące i tylko nagrania z tymi ocenami zostały wykorzystane w eksperymentach związanych z rozpoznawaniem mowy.

Wyniki lokalizacji, szczególnie w przypadku modelu *M1*, należy uznać za zadowalające. Błędy w określeniu kształtu ust bądź ich lokalizacji w przeważającej większości nagrań pojawiały się jedynie na przestrzeni kilku sąsiednich ramek; według przyjętych kryteriów takie nagranie jako całość otrzymywało wówczas niską ocenę. Gdyby oceniać każdą ramkę niezależnie, wyniki procentowe byłyby znacznie wyższe, ale

takie podejście nie byłoby odpowiednie ze względu na zagadnienie rozpoznawania mowy, w którym kształt ust w poszczególnych ramkach traktowany jest jako nierozłączna całość.

Tab. 1. Wyniki lokalizacji ust (ocena 3 – najlepsza, 0 – najgorsza)

Model M1			Model M2		
Ocena	**Liczba nagrań**	**[%] wszystkich**	**Ocena**	**Liczba nagrań**	**[%] wszystkich**
3	25	23,2	3:	22	20,4
2	57	52,8	2:	36	33,3
1	23	21,3	1:	38	35,2
0	3	2,8	0:	12	11,1
3+2	82	75,9	3+2:	58	53,7
razem	108	100	razem:	108	100

Klasyfikacja samogłosek

W roli klasyfikatora 6 polskich samogłosek (*a*, *e*, *i*, *o*, *u*, *y*) wykorzystano jednokierunkową sztuczną sieć neuronową. Na wejście sieci podawana jest macierz parametrów, czyli wektory parametrów obliczone równomiernie w 20 punktach czasu trwania wypowiedzi. W celu uniezależnienia wyników klasyfikacji od długości wypowiadanych głosek zastosowano interpolację w celu wyznaczenia wartości parametrów rozłożonych równomiernie w czasie trwania wypowiedzi. Ponadto z każdego nagrania wyznaczono 5 sekwencji przesuniętych w stosunku do siebie o wielokrotność 1/25 sekundy, co ma zapewnić sieci niewrażliwość na przesunięcie czasowe.

W procesie lokalizacji ust uzyskuje się dwa zestawy parametrów, które potencjalnie można wykorzystać jako wizualne cechy mowy ludzkiej. Pierwszym z nich jest wektor parametrów kształtu *b*, a drugim wektor parametrów luminancji bg. Wyniki klasyfikacji głosek tymi parametrami oraz parametrami *b* i b_g łącznie umieszczono w tabeli 2.

Tab. 2 Wyniki klasyfikacji sześciu polskich samogłosek z wykorzystaniem parametrów wizualnych

Model	**Liczba przykładów**	**b**		b_g		$b + b_g$	
		błędy	**[%]**	**błędy**	**%**	**błędy**	**[%]**
M1	200	137	31,50	98	51,00	110	45,00
M2	135	59	56,30	41	69,63	30	77,78

Z powyższej tabeli wynika, że wyniki rozpoznawania są wyraźnie lepsze w przypadku modelu *M2*, który wierniej oddaje kształt ust. Spośród obu grup parametrów lepsze wyniki uzyskuje się dla parametrów bg opisujących rozkład odcieni szarości wokół konturu ust, a nie dla parametrów *b* charakteryzujących sam kształt ust. Parametry b_g zawierają bowiem dodatkowe informacje o otoczeniu ust, takie jak widoczność języka i zębów, a przy tym są niezależne od ich anatomicznego kształtu. W dodatku dla modelu *M1* połączenie obu tych grup parametrów skutkuje pogorszeniem trafności klasyfikacji. Pozwala to wysnuć wniosek, że znacznie istotniejsze w procesie rozpoznawania mowy są wewnętrzne krawędzie obu warg i dodanie tej komplementarnej informacji do parametrów bg dla modelu *M2* pozwoliło uzyskać lepszy wynik, sięgający prawie 80%. Szczegółowe wyniki klasyfikacji tym modelem oraz parametrami *b* i b_g umieszczono w tabeli 3.

Tab. 3. Wyniki klasyfikacji par głosek wykorzystaniem parametrów wizualnych $b + b_g$ i modelu *M2*

Głoska	**a**	**e**	**i**	**o**	**u**	**y**	**Razem**
Liczba przykładów	15	15	25	30	30	20	135
Liczba błędów	5	10	8	2	0	5	30
[%] poprawnych	66,67	33,33	68,00	93,33	100	75,00	77,78

Skuteczność rozpoznawania mowy jest zgodna z oczekiwaniami. Najwięcej błędów zostało popełnionych w przypadku klasyfikacji tych samogłosek, których rozróżnienie tylko na podstawie analizy obrazu twarzy mówcy jest praktycznie niemożliwe. Należą do nich takie pary jak: *a – e*, *i – y*, które były szczególnie często ze sobą mylone. Przy próbach klasyfikacji tylko tych par samogłosek niezależnie uzyskano skuteczność 76,67% dla pierwszej pary oraz 95,56% dla drugiej pary. W przypadku wszystkich pozostałych par samogłosek, przy których wymowie wygląd ust jest wyraźnie odmienny, uzyskano stuprocentową skuteczność.

Dodatkowo przeprowadzono eksperymenty z rozpoznawania mowy z wykorzystaniem parametrów akustycznych. Wykorzystano w tym celu współczynniki melcepstralne MFCC (ang. *Mel Frequency Cepstral Coefficients*) [Davis, Mermelstein 1980]. Bazują one na krótkookresowym widmie sygnału. Próbki widma są grupowane i wygładzane zgodnie ze skalą melową, która odwzorowuje perceptualne własności słuchu. Następnie widmo jest dzielone na pasma krytyczne z wykorzystaniem banku filtrów. Ostatecznie dokonuje się transformacji kosinusowej logarytmów wartości na wyjściach banku filtrów, w wyniku czego uzyskuje się wektor nieskorelowanych współczynników melcesptralnych.

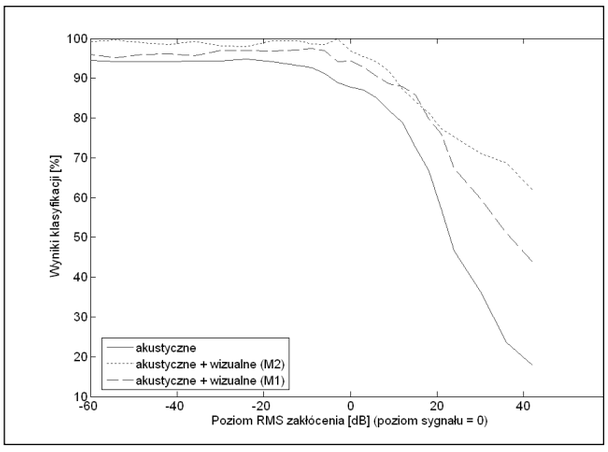
W eksperymentach wykorzystano 10 pierwszych współczynników MFCC wyznaczanych w ramkach o długości 10 ms. Wyniki klasyfikacji głosek z wykorzystaniem parametrów akustycznych i wizualnych umieszczono w tabeli 4.

Tab. 4. Wyniki klasyfikacji sześciu polskich samogłosek z wykorzystaniem parametrów wizualnych i akustycznych

Model	**Parametry akustyczne**	**Parametry wizualne i akustyczne**
M1	94,50%	96,00%
M2		99,26%

Wyniki klasyfikacji z wykorzystaniem parametrów akustycznych są znacznie wyższe, niż przy użyciu tylko parametrów wizualnych (por. tab. 2). Jednakże łączne wykorzystanie obu typów parametrów skutkuje dalszym wzrostem skuteczności rozpoznawania o 1,5 punktu procentowego w przypadku modelu *M1* i o prawie 5 punktów procentowych dla modelu *M2*.

Zbadano również odporność przedstawionego systemu na zakłócenia związane z występującym jednocześnie tłem dźwiękowym. Jako zakłócenie wykorzystano fragment mowy ludzkiej, który został dodany do klasyfikowanych nagrań. Na rycinie 2 zobrazowano zależność wyników klasyfikacji systemów opartych tylko na parametrach akustycznych oraz łącznie na parametrach akustycznych i wizualnych od wzrastającego poziomu zakłócenia.



Ryc. 2 Wyniki klasyfikacji systemów opartych o parametry wizualne i akustyczne w zależności od poziomu zakłócenia dźwięku

Skuteczność klasyfikacji systemu wykorzystującego tylko parametry akustyczne zaczyna spadać, gdy zakłócenie osiąga poziom –20 dB (dziesięć razy mniejszy niż poziom mowy). Z kolei system wykorzystujący oba typy parametrów zachowuje wysoką skuteczność dopóki poziom zakłócenia nie zrówna się z poziomem sygnału użytecznego. W przypadku bardzo dużych zakłóceń (przekraczających poziom mowy ponad sto razy) system wykorzystujący model *M2* i oba typy parametrów prawidłowo sklasyfikował znacznie więcej nagrań (62%), niż system wykorzystujący model *M1* (44%). Dowodzi to, że model *M2* niesie w sobie więcej istotnych informacji o ruchu ust i pozwala tworzyć systemy bardziej odporne na zakłócenia akustyczne.

Zakończenie

Uzyskana skuteczność rozpoznawania mowy bliska 80% na podstawie analizy twarzy mówcy dla najlepszego modelu i zestawu parametrów jest wynikiem zadowalającym, biorąc pod uwagę ogromne podobieństwo w ruchach ust towarzyszących wymowie niektórych samogłosek oraz ogromną różnorodność wynikającą z cech anatomicznych mówców i sposobu mówienia. Aby poprawić osiągane rezultaty, należy przede wszystkim udoskonalić algorytm lokalizacji ust, który w obecnej formie popełnia jeszcze wiele błędów. W tym celu niezbędne jest zwiększenie liczby mówców w bazie. Ponadto wskazane jest poszukiwanie innych parametrów, które bazując na znanym położeniu ust w obrazie pozwolą wierniej opisywać wizualną stronę mowy.

Rozpoznawanie mowy na podstawie analizy obrazu twarzy mówcy jest zadaniem znacznie bardziej skomplikowanym, niż wykorzystanie parametrów akustycznych. Jednak uzyskane wyniki potwierdzają przydatność informacji wizualnej w procesie klasyfikacji mowy.

Istnieje wiele potencjalnych zastosowań dla systemu wyznaczania i analizy wizualnych cech mowy, szczególnie w audiologii i terapii logopedycznej osób z wadami słuchu. Może on być przede wszystkim wykorzystany w trenowaniu poprawnej wymowy pacjentów z uszkodzonym słuchem.. Innym zastosowaniem jest wykrywanie, a następnie wspomaganie leczenia rozmaitych wad głosu i wymowy. Przedstawione algorytmy mogą również stanowić podstawę aplikacji zdolnej do prowadzenia interaktywnych zajęć logopedycznych w dziedzinie artykulacji głosek, przeznaczonej szczególnie dla dzieci. Opisane rozwiązanie może się sprawdzić również w osobistych systemach rozpoznawania mowy dla osób niedosłyszących, przydatnych szczególnie w hałaśliwym otoczeniu. W tym celu niezbędne będzie jednak poszerzenie bazy nagrań w fazie początkowej o wszystkie dźwięki języka polskiego, zaś na kolejnym etapie wzbogacenie bazy o mniejsze jednostki fonologiczne.

Bibliografia

Cootes T., Taylor C., Cooper D., Graham J. [1995]. Active shape Models – Their Training and Application. „Computer Vision and Image Understanding” 61(1), 38–59.

Davis S. B., Mermelstein P. [1980]. Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. „IEEE Trans. Acoustics, Speech & Signal Processing” 28(4), 357–366.

Dodd B, Campbell R. [1987]. Hearing by Eye: The Psychology of Lipreading. Lawrence Erlbaum Press.

Jackson J. [1991]. A User's Guide to Principal Components. John Wiley and Sons, Inc.

Kostek B., Dalka P. [2005]. Combining Visual and Acoustic Modalities to Ease Speech Recognition by Hearing Impaired People. 118th Audio Engineering Society Convention, Paper No. 6462, Barcelona, Spain.

Luettin J., Thacker N. [1997]. Speechreading Using Probabilistic Models. „Computer Vision and Image Understanding” 65(2), 163–178.

Nelder J., Mead R. [1965]. A simplex method for function optimization. „Computing Journal” 7(4), 308 313.

Summerfield Q. [1992]. Lipreading and Audio-Visual Speech Perception. „Philosophical transactions of the Royal Society of London. Series B, Biological sciences” 335, 71–78.

Podziękowania

Badania były dofinansowane przez Fundację na Rzecz Nauki Polskiej.

Adres do koresponencji

mgr inż. Piotr Dalka
Katedra Systemów Multimedialnych Politechniki Gdańskiej
ul. Narutowicza 11/12
80-952 Gdańsk
e-mail: dalken@sound.eti.pg.gda.pl