STUDIA INFORMATICA 2013
Volume 34 Number 2B (112)

Dorota KAMIŃSKA, Tomasz SAPIŃSKI, Dominik NIEWIADOMY, Adam PELIKANT
Politechnika Łódzka, Instytut Mechatroniki i Systemów Informatycznych

PORÓWNANIE WYDAJNOŚCI WSPÓŁCZYNNIKÓW PERCEPTUALNYCH NA POTRZEBY AUTOMATYCZNEGO ROZPOZNAWANIA EMOCJI W SYGNALE MOWY

Streszczenie. Przedmiotem niniejszego artykułu jest parametryzacja sygnału mowy emocjonalnej przy użyciu współczynników preceptualnych. Dokonano porównania wydajności współczynników MFCC z współczynnikami HFCC oraz przynależnych im parametrów dynamicznych. Na podstawie bazy mowy emocjonalnej oceniono skuteczność wybranych współczynników.

Słowa kluczowe: rozpoznawanie emocji, sygnał mowy, MFCC, HFCC

COMPARISON OF PERCEPTUAL FEATURES EFFICIENCY FOR AUTOMATIC IDENTIFICATION OF EMOTIONAL STATES FROM SPEECH SIGNAL

Summary. The following paper presents parameterization of emotional speech using perceptual coefficients. The comparison of MFCC to HFCC and adherent dynamic parameters is presented. Basing on emotional speech database efficiency of used coefficients was evaluated.

Keywords: emotions recognition, speech signal, MFCC, HFCC

1. Wprowadzenie

Komunikacja interpersonalna to nieodzowny elementem ludzkiego życia. Rozmowa dostarcza słuchaczowi zarówno informacji lingwistycznych, jak i nakreśla charakterystykę biologiczno-psychologiczną mówcy. Posiadanie takich informacji polepsza jakość komunikacji. Ważnym elementem konwersacji jest ocena stanu emocjonalnego rozmówcy. Emocje wyrażane są poprzez procesy zarówno werbalne (intonacja), jak i niewerbalne (m.in. gesty, mimi-

ka twarzy, kontakt wzrokowy). W dzisiejszych czasach, kiedy komputery są częścią naszego życia, poszukuje się rozwiązań mających na celu polepszenie komunikacji człowiek-komputer/człowiek-robot (HCI/HRI), dlatego też powstają nowoczesne technologie rozpoznawania ludzkiej mowy. Systemy, które dodatkowo rozpoznawałyby stany emocjonalne użytkownika, stałyby się bardziej naturalne i wiarygodne, też komputerowe rozpoznawanie emocji stało się istotnym trendem badawczym [1].

Zaprezentowane w artykule rozważania dotyczą ekspresji emocji zawartych w sygnale mowy. Podczas wypowiedzi sygnał akustyczny generowany jest przez struny głosowe, a następnie modyfikowany przez trakt głosowy (usta, język, jama nosowa), dlatego też wpływ na odbiór stanu emocjonalnego mają takie parametry, jak prozodia, częstotliwość fałdów głosowych czy częstotliwości formantów. Badania wykazują, że wykorzystanie parametrów perceptualnych MFCC, używanych tak szeroko w zagadnieniach analizy sygnału mowy, daje wysoką jakość klasyfikacji także w omawianym zagadnieniu [2]. Wysoka skuteczność rozpoznawania mowy na ich podstawie wiąże się z podejściem naśladowania postrzegania dźwięku przez ludzkie ucho. Dlatego też w niniejszych badaniach autorzy skupili się na sprawdzeniu jakości klasyfikacji innych współczynników perceptualnych, wykazujących dobre rezultaty w rozpoznawaniu mowy, a niewykorzystywanych w rozpoznawaniu mowy emocjonalnej.

Dalsza część niniejszej pracy została podzielona na cztery rozdziały. Rozdział drugi prezentuje bazy mowy emocjonalnej wykorzystane w niniejszych badaniach. Rozdział trzeci prezentuje przegląd parametrów preceptualnych wykorzystanych do opisu sygnału mowy. W rozdziale czwartym zostały przedstawione wyniki klasyfikacji emocji oraz skuteczność wybranych parametrów. Rozdział piąty stanowi krótkie podsumowanie wykonanych badań oraz przyszłe kierunki rozwoju.

2. Materiał badawczy

Na potrzeby niniejszych badań wykorzystano dwie bazy mowy emocjonalnej. Pierwszą z nich stanowi polska baza emocji, udostępniana przez Zakład Elektroniki Medycznej Politechniki Łódzkiej. Stanowi ona zbiór 240 nagrań pięciu różnych zdań wypowiadanych przez ośmiu aktorów (4 kobiety, 4 mężczyzn). Ze względu na zabarwienie emocjonalne, zbiór podzielono na grupy (klasy): radość (H), smutek (S), złość (A), strach (F), znudzenie (B) oraz mowa neutralna (N). Jakość nagrań została oceniona przez pięćdziesięciu decydentów, którzy dokonywali klasyfikacji losowo wybranych próbek z każdej grupy emocji. Średni poziom rozpoznawania wśród decydentów wynosił 72% [3].

Drugą bazę stanowią również próbki mowy odegranej, aczkolwiek w języku niemieckim – Berlin Emotional Speech Database jest najczęściej wykorzystywaną bazą w tego typu

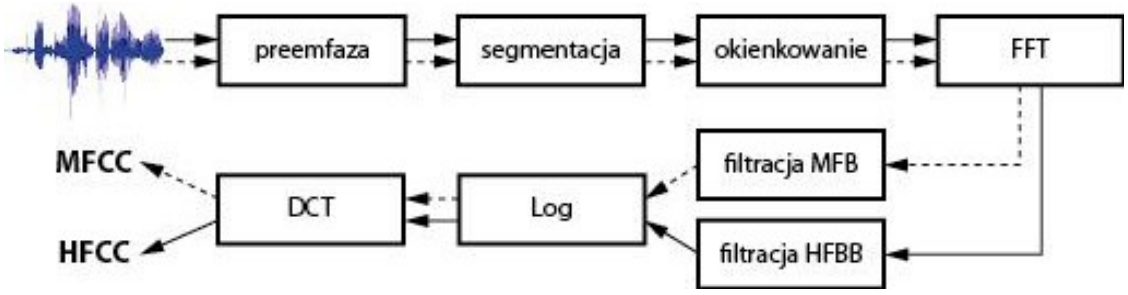
badaniach. Zawiera ona 525 nagrań pochodzących od 10 mówców (5 mężczyzn, 5 kobiet). Próbki sklasyfikowano na testowej grupie słuchaczy, przypisując im 7 stanów emocjonalnych ze zbioru: radość (H), smutek (S), złość (A), strach (F), znudzenie (B), obrzydzenie (D), mowa neutralna (N) [4].

3. Parametryzacja sygnału mowy

Reprezentacja sygnału mowy w dziedzinach czasu i częstotliwości jest zbyt złożona, by stanowić dane wejściowe klasyfikatora, dlatego poszukuje się parametrów odzwierciedlających opis ilościowy sygnału. Najczęściej wykorzystywaną grupą deskryptorów opisujących mowę są charakterystyki częstotliwościowe oraz spektralne, w tym charakterystyka częstotliwości podstawowej tonu krtaniowego i częstotliwości formantowe [5]. Szeroko stosowane są również współczynniki predykcji liniowej LPC [6] oraz współczynniki wykorzystujące podejście perceptualne, które naśladuje mechanizmy rozpoznawania sygnału mowy, występujące u człowieka – współczynniki MFCC [7] .

Podejście perceptualne polega na przekształceniu częstotliwości tak, aby odpowiadała subiektywnemu odbiorowi przez ludzki aparat słuchowy. W tym celu stosuje się skale perceptualne, np. Mel czy Bark. W związku z badaniami [8], postanowiono porównać wpływ współczynników MFCC oraz HFCC na rozpoznawanie mowy pod kątem zabarwienia emocjonalnego. Należy pamiętać, że parametry te nie stanowią całego wektora cech opisujących sygnał, a tylko jego część, natomiast badania te stanowią jedynie porównanie wyżej wymienionych współczynników.

W tym rozdziale zaprezentowano wykorzystane parametry perceptualne (MFCC, ΔMFCC, ΔΔMFCC, HFCC, ΔHFCC, ΔΔHFCC). Schemat ekstrakcji parametrów prezentuje rys. 1.



Rys. 1. Algorytm ekstrakcji parametrów MFCC i HFCC
Fig. 1. MFCC and HFCC extraction algorithm

3.1. Współczynniki MFCC

Współczynniki Mel Frequency Cepstral Coefficients (MFCC) znajdują szerokie zastosowane w dziedzinie analizy mowy (systemy rozpoznawania mówcy, mowy czy emocji). Spo-

wodowane jest to zastosowaniem podejścia perceptualnego, odwzorowującego percepcje przez ludzki narząd słuchu.

Wstępną obróbkę sygnału stanowi proces preemfazy, w efekcie którego wzmacniane są wysokie częstotliwości sygnału, co skutkuje dużą odpornością na zakłócenia otoczenia. Następnie dokonywana jest segmentacja sygnału na ramki o długości 25 ms, z przesunięciem wynoszącym 50% szerokości ramki. Podzielony w ten sposób sygnał poddawany jest operacji iloczynu z oknem Hamminga. Następnie dla każdej ramki wykonywane jest obliczenie mocy widma przy użyciu szybkiej transformaty Fouriera FFT. Kolejnym krokiem jest nałożenie na otrzymany wynik trójkątnych filtrów z banku filtrów uwzględniających skalę Mel (MFB). Skala mel jest skalą logarytmiczną, naśladującą sposób, w jaki ludzkie ucho postrzega dźwięk. Poniższe równania prezentują tę samą częstotliwość w skali Mel (1) i skali Hertz (2):

$$\mathrm{m} = 1127,01048 \ln\left(1 + \frac{\mathrm{f}}{700}\right) \tag{1}$$

$$\mathrm{f} = 700\left(\mathrm{e}^{\frac{\mathrm{m}}{1127,01048}} - 1\right) \tag{2}$$

Ostatnim krokiem algorytmu jest obliczenie dyskretnej transformaty kosinusowej DCT na wyniku otrzymanym z poprzedniego etapu przetwarzania, poddanym logarytmizacji według poniższego równania:

$$\mathrm{c_k} = \sqrt{\frac{2}{L}} \sum_{l=0}^{L-1} \ln \tilde{S}(l) \cos\left(\frac{\Pi k}{L}\left(L + \frac{1}{2}\right)\right) \tag{3}$$

gdzie: k=0,1 ,…,q−1, q – liczba współczynników, L – liczba zastosowanych filtrów

Następnie, na podstawie współczynników MFCC, wyznaczono parametry dynamiczne. Parametry będące pochodną współczynników, po czasie wskazują na prędkość zmian sygnału i definiowane są wzorem (3):

$$\Delta\mathrm{MFCC_n}(c) = \frac{\sum_{i=-\alpha}^{\alpha} i \cdot MFCC_{n+i}(c)}{2\sum_{i=-\alpha}^{\alpha} i^2} \tag{4}$$

Parametry będące pochodną parametrów ΔMFCC, opisują przyspieszenie zmian sygnału i definiowane są wzorem (4):

$$\Delta\Delta\mathrm{MFCC_n}(c) = \frac{\sum_{i=-\alpha}^{\beta} i \cdot \Delta MFCC_{n+i}(c)}{2\sum_{i=-\alpha}^{\beta} i^2} \tag{5}$$

gdzie: współczynniki α (w przypadku ΔMFCC) i β (w przypadku ΔΔMFCC) decydują o szerokości spektrum ramek użytych podczas liczenia pochodnych.

3.2. Współczynniki HFCC

Analiza zaproponowana przez Marka Skowrońskiego i Johna Harrisa [9] wykazuje większe podobieństwo do percepcji mowy, dzięki czemu pozwala uzyskać lepsze rezultaty w rozpoznawaniu [8]. Poprzez wprowadzenie niewielkiej modyfikacji algorytmu generacji parametrów MFCC, wyznaczane są współczynniki HFCC. Modyfikacja ta polega jedynie na zmianie banku filtrów Melowych (MFB) na nowy zestaw filtrów (HFBB), w którym filtry są szersze oraz w większym stopniu nachodzą na siebie. Sprowadza się to do podmiany jednej części algorytmu i nie wpływa na jego wydajność. Metodę generacji banku filtrów HFCC szerzej opisano w [8]. Podobnie jak w przypadku współczynników MFCC, dla współczynników HFCC wyznaczono współczynniki dynamiczne ΔHFCC oraz ΔΔHFCC.

4. Eksperymenty

Na podstawie opisanych w rozdziale trzecim algorytmów, zostały stworzone oddzielne wektory cech dla każdej z wypowiedzi zawartych w dwóch wykorzystanych w badaniach bazach. Rodzaj i liczba parametrów w wektorach cech uzależnione są od przeprowadzonych badań, opisanych w poniższych podrozdziałach.

Klasyfikacji dokonano przy użyciu algorytmu kNN, o wartości współczynnika $k = 5$. Wartość ta została dobrana doświadczalnie, jako dająca najwyższe rezultaty klasyfikacji. Do obliczania odległości wykorzystano standardową metrykę euklidesową.

4.1. Wpływ liczby parametrów na jakość rozpoznawania

Tabela 1

Wpływ liczby współczynników MFCC i HFCC na jakość rozpoznawania emocji

Liczba parametrów	Skuteczność rozpoznawania [%]	
	MFCC	HFCC
11	55,92	56,28
12	56,64	58,05
13	**58,**7	**59,29**
14	57,87	57,34

W wyniku parametryzacji dla każdej ramki sygnału wyznaczono współczynniki MFCC oraz HFCC. Dla każdego bloku z uzyskanego zbioru współczynników wyznaczono wiele opisujących go cech statystycznych: średnia, mediana, odchylenie standardowe, minimum i maksimum. Na podstawie tych cech stworzono wektory, które następnie poddano klasyfikacji. W ramach badań zweryfikowano wpływ liczby współczynników MFCC oraz HFCC na

jakość rozpoznawania. W ten sposób ustalono optymalny rozmiar wektora cech, używany w poniższych badaniach. Wyniki przestawiono w tabeli 1.

4.2. Wpływ parametrów dynamicznych na jakość rozpoznawania emocji

Do zestawów cech opisanych w podrozdziale 4.1 dodano kolejno parametry dynamiczne ΔMFCC, ΔHFCC, ΔΔMFCC, ΔΔHFCC, opisane w rozdziale 3. Skuteczność klasyfikacji z różnymi zestawieniami cech przedstawia tabela 2.

Tabela 2

Wpływ parametrów dynamicznych na jakość rozpoznawania emocji

	Skuteczność rozpoznawania [%]	
Parametry	Baza niemiecka	Baza polska
MFCC	58,9	50,73
MFCC + ΔMFCC	60	50,39
MFCC + ΔΔMFCC	60,35	**51,97**
MFCC + ΔMFCC + ΔΔMFCC	**60,53**	51,29
HFCC	59,29	53,1
HFCC + ΔHFCC	60,53	**55,48**
HFCC + ΔΔHFCC	**61,59**	54,46
HFCC + ΔHFCC + ΔΔHFCC	60,17	53,67

5. Podsumowanie

Pierwszy etap badań stanowi analiza wpływu liczby współczynników MFCC i HFCC na jakość rozpoznawania. Biorąc pod uwagę fakt, iż w większości publikacji dotyczących sygnału mowy liczba współczynników MFCC wynosi 12 lub 13, sprawdzona została ta właśnie liczebność wraz z wartościami sąsiadującymi. Osiągnięte rezultaty, zaprezentowane w tabeli 1, potwierdzają, że 13 to liczba współczynników dająca najlepsze wyniki rozpoznawania, zarówno dla MFCC, jak i dla HFCC. Bazując na uzyskanych wynikach, resztę badań przeprowadzono z użyciem 13 parametrów.

Kolejnym krokiem badań są porównanie skuteczności opisanych wyżej parametrów oraz ocena wpływu współczynników dynamicznych na jakość rozpoznawania. Zarówno dla bazy polskiej, jak i bazy niemieckiej lepszą jakość rozpoznawania osiągnięto przez zastosowanie współczynników HFCC. Wykorzystanie owych parametrów nie niesie ze sobą dodatkowych kosztów obliczeniowych w porównaniu z parametrami MFCC, natomiast daje widocznie lepsze rezultaty w rozpoznawaniu mowy emocjonalnej. Jak widać w tabeli 2, współczynniki dynamiczne polepszają jakość rozpoznawania dla obu typów parametrów. Należy zaznaczyć jednak, iż wiąże się to ze wzrostem rozmiaru przestrzeni cech, co skutkuje wzrostem kosztu obliczeniowego procesu klasyfikacji. Należy również zauważyć wpływ próbek mowy na ja-

kość rozpoznawania. Jak widać w powyżej zaprezentowanych tabelach, skuteczność klasyfikacji dla bazy niemieckiej znacznie przewyższa skuteczność uzyskaną dla bazy polskiej. Wpływ ten jest w szczególności zauważalny w badaniach przedstawionych w [10], gdzie dla samych współczynników MFCC uzyskano wyniki powyżej 80%.

Kolejnym etapem badań powinno być przeprowadzenie porównania jakości rozpoznawania dla współczynników HFCC i MFCC, z użyciem innych baz mowy emocjonalnej, w celu potwierdzenia zaprezentowanych obserwacji, tj. wyższej skuteczności rozpoznawania z wykorzystaniem współczynników HFCC. Należałoby też potwierdzić skuteczność owych parametrów dla mowy spontanicznej, w której to próbki nie są jednoznacznie określone, natomiast mogą być mieszaniną kilku emocji podstawowych jednocześnie.

BIBLIOGRAFIA

1. Ślot K.: Rozpoznawanie biometryczne. Wydawnictwa Komunikacji i Łączności, Warszawa 2010.
2. Tu M.-C., Liao W.-K., Chin Y.-H., Lin C.-H., Liao W.-J., Lin S.-H., Wang J.-C.: Speech Based Boredom Verification Approach for Modem Education System. International Symposium on Information Technology in Medicine and Education (ITME), 2012.
3. Cichosz J.: Database of Polish Emotional Speech. http://www.eletel.p. lodz.pl/med/eng/.
4. Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B.: A Database of German Emotional Speech. Proc. Interspeech 2005.
5. Narayanan S., Busso C., Lee S.: Analysis of emotionally salient aspects of fundamental frequency for emotion detection. Proc. IEEE Transactions on Audio, Speech, AND Language Processing, No. 17, 2009.
6. Pathak S., Kulkarni A.: Recognizing emotions from speech. Proc. 3rd International Conference on Electronics Computer Technology (ICECT), Vol. 4, 2011.
7. Bedoya-Jaramillo S., Belalcazar-Bolaños E., Villa-Cañas T., Orozco-Arroyave J. R., Arias-Londoño J. D., Vargas-Bonilla J. F.: Automatic Emotion Detection in Speech Using Mel frequency Cesptral Coefficients. Proc. 2012 XVII Symposium of Image, Signal Processing, and Artificial Vision (STSIVA), 2012.
8. Niewiadomy D.: Detekcja izolowanych słów w nagraniach dla potrzeb wdrożenia mechanizmu automatycznych wyzwalaczy audio w systemach baz danych. Politechnika Łódzka, Praca doktorska, 2012.
9. Skowroński M., Harris J.: Increased mfcc filter bandwidth for noise-robust phoneme recognition. ICASSP 2002, s. 801÷804.

10. Murali Krishna N., Lakshmi P. V., Srinivas Y., Sirisha Devi J.: Emotion Recognition using Dynamic Time Warping Technique for Isolated Words. IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No. 1, 2011.

Wpłynęło do Redakcji 16 stycznia 2013 r.

Abstract

Following paper presents parameterization of emotional speech using perceptual coefficients as well as a comparison of Mel Frequency Cepstral Coefficients, Human Factor Cepstral Coefficients and adherent dynamic parameters effect on emotion recognition. Analysis was performed on two different databases: Database of Polish Emotional Speech and the most commonly used for emotion recognition – Berlin Database of Emotional Speech. Both consist of acted emotional speech grouped into seven classes of primary emotions.

The first step of the analysis was MFCC and HFCC feature extraction from speech signal, the algorithm of which is presented in Figure 1. Next, the effect of the number of perceptual parameters was investigated. The results are presented in Table 1, basing on them further analysis was performed using 13 HFCC an MFCC parameters. Subsequently, the effect of dynamic parameters on accuracy performance was analyzed and presented in Table 2.

Application of HFCC does not pose additional computational cost compared to the MFCC parameters, and gives visibly better results for emotion recognition from speech. Moreover, it was noticed, that dynamic parameters improve recognition quality for both types of coefficient (MFCC and HFCC). However, it should be noted, that adding them to the feature vector, increases computing cost of classification.

Adresy

Dorota KAMIŃSKA: Politechnika Łódzka, Instytut Mechatroniki i Systemów Informatycznych, ul. Stefanowskiego 18/22, 90-924 Łódź, Polska, dorota.kaminska@p.lodz.pl.

Tomasz SAPIŃSKI: Politechnika Łódzka, Instytut Mechatroniki i Systemów Informatycznych, ul. Stefanowskiego 18/22, 90-924 Łódź, Polska.

Dominik NIEWIADOMY: Politechnika Łódzka, Instytut Mechatroniki i Systemów Informatycznych, ul. Stefanowskiego 18/22, 90-924 Łódź, Polska.

Adam PELIKANT: Politechnika Łódzka, Instytut Mechatroniki i Systemów Informatycznych, ul. Stefanowskiego 18/22, 90-924 Łódź, Polska.