Piotr Żelasko, Agata Trawińska, Bartosz Ziółko,
Marcin Czyżyk, Joanna Stanisławek, Elżbieta Ślusarz

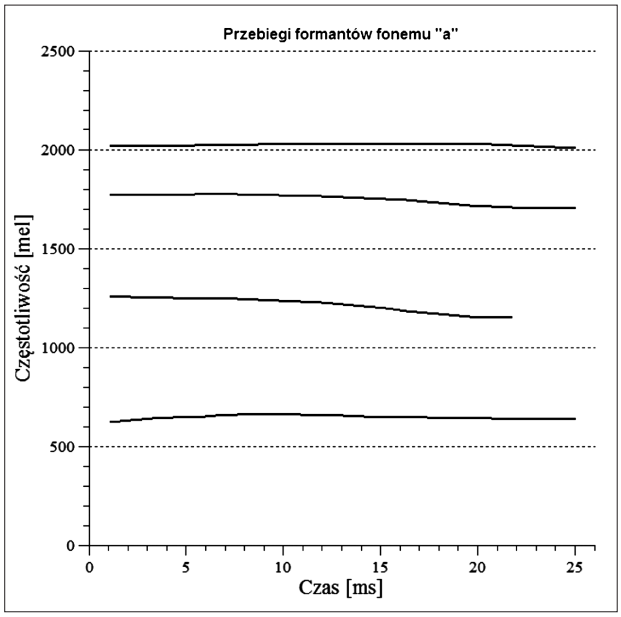
Zastosowanie algorytmu DTW jako narzędzia w identyfikacji mówcy

Wstęp

W badaniach fonoskopijnych dotyczących identyfikacji mówców wypracowane zostały różne metodyki badań, w tym najczęściej używana w polskiej praktyce sądowej, tzw. metoda językowo-pomiarowa [2], [3]. Uwzględnia ona zarówno analizę nawyków artykulacyjnych mówiącego, jak i analizę parametrów akustycznych mowy – najczęściej struktur formantowych wyekstrahowanych z segmentów samogłoskowych. Analiza nawyków artykulacyjnych ma charakter głównie jakościowy i dokumentowana jest transkrypcją fonetyczną dystynktywnych wymówień. Z kolei analiza akustyczna, niezależnie od używanych parametrów, np. częstotliwość tonu krtaniowego F0, częstotliwości kolejnych formantów F1, F2, F3, F4, współczynniki cepstralne, energia sygnału [11], jako oparta na danych ilościowych jest przedmiotem różnego rodzaju statystycznych opracowań zmierzających do efektywnego i ekonomicznego wykorzystania informacji zawartych w sygnale mowy. Przykładowo, w metodzie językowo-pomiarowej stosowanej w IES *gros* analizy akustycznej to ekstrakcja częstotliwości formatowych, które mają relatywnie bezpośrednie korelaty artykulacyjne i stąd preferowane są przez językoznawców. Najczęściej, z uwagi na ograniczenia pasma analizowanych nagrań, możliwa jest ekstrakcja czterech najniższych formantów F1–F4.

W obszarze identyfikacji mówców używa się także metodologii zupełnie odmiennej niż tradycyjne podejście lingwistyczne. Wraz z rozwojem technik cyfrowych i nauk informatycznych dynamicznie rozwinęły się prace nad systemami automatycznego rozpoznawania mowy, a wypracowane w nich rozwiązania przeniesione zostały do identyfikacji mówców. Systemy takie dokonują parametryzacji sygnału mowy, segmentując sygnał na małe fragmenty (okna czasowe), a następnie przekształcając zawarte w ramce dane tak, by uzyskać reprezentację o chwilowej konfiguracji traktu głosowego. Do takiego opisu używa się często melowych współczynników cepstralnych (Mel Frequency Cepstral Coefficients – MFCC) [7] lub kodowania liniową predykcją (Linear Prediction Coding – LPC) [8]. W dalszej kolejności przykładowy system może w sposób automatyczny dokonać porównania wcześniej sparametryzowanej ramki z bazą danych populacyjnych, gdzie zawarte są informacje o innych mówcach. Jednym ze sposobów na dokonanie takiego porównania jest zastosowanie ukrytych modeli Markowa (Hidden Markov Model – HMM) [1], [9], które pozwalają na oszacowanie prawdopodobieństwa, że konkretny mówca wypowiedział analizowaną część kwestii. Wskutek tego system podejmuje decyzję o zakwalifikowaniu nieznanego mówcy jako tego, który maksymalizuje prawdopodobieństwo wypowiedzenia badanego sygnału mowy [4], [6]. Przedstawiona tu przykładowa procedura bliższa jest weryfikacji, czyli temu, z czym można się spotkać w systemach bankowych, w których posiadacz lub krąg posiadaczy pewnych uprawnień jest znany i ściśle ograniczony, niż właściwej identyfikacji mówcy, której zbiór osób, spośród których należy zidentyfikować mówcę, jest otwarty; systemy te wykazują się jednak użytecznością przy rozpoznawaniu mówców [9].

Przedstawione w artykule badania są powiązane z pracą inżynierską głównego autora. Stwierdzono w nich



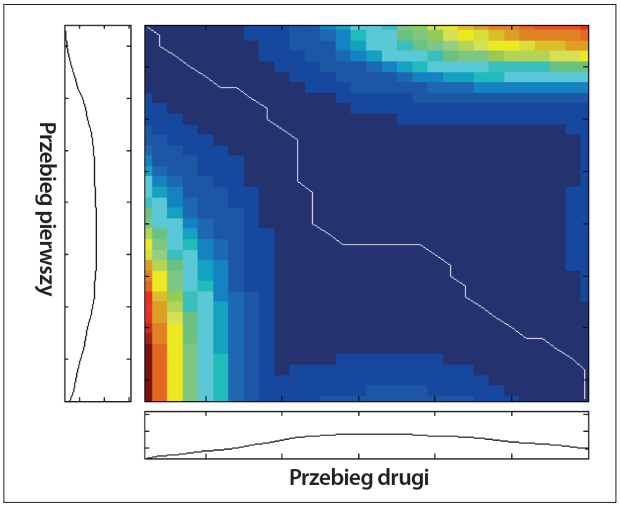
Ryc. 1. Czasowy przebieg formantów F1, F2, F3 i F4 (rosnąco wraz częstotliwością) dla samogłoski „a" z kontekstu „pas"
Fig. 1 *Time course of F1, F2, F3 and F4 formants (ascending with frequency) for "a" vowel in "pas" context*
Źródło: autorzy

m. in., że im wyższy jest formant, tym bardziej różnicuje on mówców, a wyłączenie F1 z analizy nie pogarsza w znaczącym stopniu rozpoznawalności [13].

Algorytm DTW

Przedmiotem badań opisanych w pracy jest możliwość zastosowania algorytmu DTW jako narzędzia przyspieszającego proces identyfikacji nieznanego mówcy w zakresie danych akustycznych i przy założeniu adekwatności fonetycznej wypowiedzi dowodowych i referencyjnych [1], [5]. Standardowe zastosowanie algorytmu to wykorzystanie jego własności dopasowania do siebie dwóch przebiegów czasowych, które są rozsynchronizowane względem siebie lub różnią się od siebie szybkością przebiegu. Takimi przebiegami mogą być np. sparametryzowane sygnały samogłosek jednego mówcy, w przypadku gdy za pierwszym razem mówi wolniej, a za drugim szybciej lub gdy w odmiennych warunkach komunikacyjnych zastosował inne procesy językowe. DTW jest też stosowany np. w dopasowaniu do siebie odpowiednio sparametryzowanych podpisów (przeniesionych do domeny cyfrowej) lub dopasowaniu sekwencji genetycznych [12].

W dużym uproszczeniu sposób działania DTW można opisać następująco: dysponując dwoma sygnałami x i y o długościach, odpowiednio N i M, algorytm układa siatkę o wymiarach N x M, gdzie w każdym jej polu obliczana jest odległość pomiędzy próbkami sygnałów o numerach n i m, gdzie n = 1, 2, …, N i m = 1, 2, …, M. Następnie poszukiwana jest ścieżka W łącząca punkty (1, 1) i (N, M) w taki sposób, aby koszt przejścia między tymi punktami był jak najmniejszy. Ścieżka ta musi spełniać warunki ciągłości i monotoniczności – w każdym kroku następny punkt ścieżki musi być określony tak, aby sąsiadował z punktem poprzednim, a indeksy n i m nie mogą maleć. Tak dobrana ścieżka nazywa się optymalną ścieżką dopasowania.



Ryc. 2. Macierz kosztów przejścia z optymalną ścieżką dopasowania DTW. Cieplejsze kolory oznaczają większy koszt przejścia niż kolory chłodniejsze
Fig. 3 *Transition matrix with optimal DTW matching track. Warm shades indicate higher cost of transition than cold shades*
Źródło: autorzy

Dopasowanie do siebie dwóch sygnałów tą metodą pozwala na określenie odległości między nimi. W tym wypadku jest ona liczona jako suma wartości wszystkich pól, przez które przechodzi ścieżka, znormalizowana względem długości ścieżki. Dysponując zatem dla różnych mówców przebiegami formantowymi różnych samogłosek i zarazem poszczególnych samogłosek z różnych kontekstów fonetycznych, możemy porównywać badany sygnał z sygnałami referencyjnymi i klasyfikować go jako sygnał tego mówcy, do którego odległość była najmniejsza.

W praktyce zaproponowana procedura zakłada obliczenie odległości między wszystkimi sygnałami referencyjnymi (pochodzącymi z nagrań znanego mówcy), a następnie ustalenie osobnego progu rozpoznania dla każdego formantu F1, F2, F3, F4 jako największej odległości występującej w ramach wypowiedzi tego samego mówcy. W dalszej kolejności liczone są odległości między sygnałem testowym (pochodzącym od nieznanego mówcy) a sygnałami referencyjnymi z uwzględnieniem progów klasyfikacji wyliczonych w poprzednim etapie. Jeżeli każdy formant spełni warunek znalezienia się poniżej progu, następuje klasyfikacja mówcy. Wadą tej metody jest brak możliwości klasyfikacji w wypadku, gdy dysponujemy tylko jednym sygnałem referencyjnym – nie można wtedy wyznaczyć progu rozpoznania.

Testowanie procedury

Opisana w publikacji procedura, mająca na celu ułatwienie oraz przyspieszenie procesu identyfikacji mówcy przez selekcję dostatecznych danych akustycznych, testowana jest za pomocą nagrań bezpośrednich wypowiedzi, zarejestrowanych przy użyciu mikrofonu kierunkowego w zaadaptowanym akustycznie pomieszczeniu i zapisanych w formacie WAVE PCM (44,1kHz, 16 bitów). Dane wejściowe to czasowe przebiegi formantów F1, F2, F3 i F4 wyekstrahowane dla samogłosek {a, o, u, e, y, i} występujących w różnych kontekstach fonetycznych.

Procedura została przetestowana z użyciem danych wyekstrahowanych z wypowiedzi ośmiu mówców: w sumie uwzględniono 83 warianty realizacyjne sześciu polskich samogłosek występujących w różnych kontekstach, tj. w otoczeniu różnych głosek poprzedzających i następujących, z których każda występowała w dwóch lub trzech powtórzeniach. Weryfikacja poprawności rozpoznania przez algorytm została dokonana przez zebranie dla poszczególnych samogłosek wariantów wyekstrahowanych z takich samych kontekstów i obliczenie odległości DTW pomiędzy nimi, a następnie po kolei odnoszenie ich do progów właściwych dla każdego mówcy. Podczas takiego

testowania system rozpoznający może zaklasyfikować mówców poprawnie na dwa sposoby: zaakceptować przebieg należący do tego samego mówcy (*true acceptance* – TA) lub odrzucić przebieg należący do innego mówcy (*true rejection* – TR) oraz popełnić dwa rodzaje błędów: zaakceptować przebieg należący do innego mówcy (f*alse acceptance* – FA) lub odrzucić przebieg należący do tego samego mówcy (*false rejection* – FR). Miarą rzetelności systemu są parametry: *Precision*, *Recall* oraz *F*, zdefiniowane według podanych poniżej wzorów:

$$Precision = \frac{TA}{TA + FA}$$

$$Recall = \frac{TA}{TA + FR}$$

$$F = 2\,\frac{Precision * Recall}{Precision + Recall}$$

w których *Precision* określa odsetek poprawnie zaakceptowanych przebiegów pośród wszystkich zaakceptowanych przebiegów, *Recall* określa odsetek poprawnie zaakceptowanych przebiegów wśród przebiegów, które powinny zostać zaakceptowane, a *F* jest średnią harmoniczną powyższych parametrów i pełni funkcję liczbowego wskaźnika opisującego jakość procedury.

Tabela 1 przedstawia liczby poprawnie i niepoprawnie zakwalifikowanych przebiegów dla całego zbioru danych, a także wskaźniki jakości procedury. *Precision* wyniosło 55,3%, co oznacza, że ze wszystkich zaakceptowanych przebiegów nieco ponad połowa została zaakceptowana poprawnie. *Recall* wyniosło 100%, więc żaden przebieg, który powinien zostać zakwalifikowany, nie został odrzucony.

Wyniki przedstawione w tabeli 2 to rezultat testowania procedury wyłącznie dla wariantów samogłosek {a, e, y}. Ich ograniczony zbiór został wybrany według kryterium łatwości uzyskiwania parametrów, co z kolei wynika z frekwencyjności i struktury akustycznej tych głosek. W tym wypadku *Precision* wzrasta do 78%, a *Recall* pozostaje na poziomie 100%. Ogólna miara jakości procedury *F* wzrasta z 0,71 do 0,87.

Tabela 3 przedstawia wyniki, jakie osiągnęły poszczególne samogłoski, gdy proces identyfikacji został oparty wyłącznie na jednej z nich. Samogłoski {y, a, e} uzyskały najlepsze rezultaty, osiągając parametr *F* na poziomach kolejno 0,91, 0,9 i 0,83. Identyfikacja oparta na samogłoskach {u, i, o} jest mniej precyzyjna, na co wskazuje osiągana przez nie wartość parametru *F* – jest to kolejno 0,76, 0,72 i 0,59. Najbardziej charakterystyczne są zatem samogłoski *y* oraz *a*, a najmniej – samogłoska *o*.

Podsumowanie

Przedstawione powyżej wyniki to rezultaty wstępnych analiz przydatności algorytmu do zastosowań praktycznych. Zauważalny znaczny wzrost parametru *Precision* przy użyciu części danych potwierdza słuszność proponowanego przez biegłych wyboru przebiegów czasowych wariantów wymawianiowych trzech fonemów samogłosek {a, e, y} w środku jako efektywnych w identyfikacji mówcy. Parametr *Recall* niezmiennie przyjmuje wartość 100%, ponieważ próg kwalifikacji jest dobrany tak, by samogłoski wyekstrahowane z wypowiedzi tego samego mówcy zawsze zostały zakwalifikowane. W ramach pracy dyplomowej [13] przeprowadzone zostały ponadto badania na większej liczbie danych – pozwoliły one na określenie przydatności poszczególnych formantów przy identyfikacji mówcy.

Tab. 1.

Wyniki weryfikacji procedury pod kątem rozpoznania mówców dla określonych przez konteksty wariantów samogłosek {a, e, y, i, o, u} (83 warianty)

Results of procedure verification in terms of speaker recognition for context-selected vowels {a, e, y, i, o, u} (83 variants)

TA	TR	FA	FR	Precision	Recall	F
5547	96465	4455	0	0,55	1	0,71

Tab. 2.

Wyniki weryfikacji procedury pod kątem rozpoznania mówców dla określonych przez konteksty wariantów samogłosek {a, e, y} (47 wariantów)

Results of procedure verification in terms of speaker recognition for context-selected vowels {a, e, y} (47 variants)

TA	TR	FA	FR	Precision	Recall	F
2325	48699	657	0	0,78	1	0,88

Tab. 3.

Wyniki weryfikacji procedury pod kątem rozpoznania mówców przy wyborze pojedynczych samogłosek jako podstawy do identyfikacji

Results of procedure verification in terms of speaker recognition with selection of single vowels as basis for identification

Samogłoska	TA	TR	FA	FR	Precision	Recall	F
a	835	18151	177	0	0,83	1	0,90
o	2327	26694	3202	0	0,42	1	0,59
u	648	14538	408	0	0,61	1	0,76
e	892	16416	356	0	0,71	1	0,83
y	598	13768	124	0	0,83	1	0,91
i	247	6534	188	0	0,57	1	0,72

Otrzymane statystyki rozpoznania wydają się niewystarczająco dobre, by w tej formie zastosować DTW wprost do procesu identyfikacji mówców, a właściwie jego części opartej na danych akustycznych. Możliwe, że zaimplementowanie bardziej złożonej wersji algorytmu (ograniczenia ścieżki, inna metryka, itd.), pozwoliłoby uzyskać lepsze rezultaty. Na podstawie zaprezentowanych wyników można jednak wnosić, że zaproponowany algorytm jest użyteczny do ilościowego określenia, które samogłoski i formanty niosą więcej informacji o indywidualizujących nawykach wymówieniowych osoby. Dotychczas przeprowadzone badania pozwalają na jednoczesne zmniejszenie ilości danych, jakie biegły musi poddać, zwykle półautomatycznej, a więc czasochłonnej, analizie akustycznej, oraz zwiększenie precyzji identyfikacji mówcy przez odrzucenie części danych, mało charakterystycznych dla mówców.

Część prac była finansowana przez MNiSW w ramach działalności statutowej AGH.

BIBLIOGRAFIA

1. L.R. Rabiner, R.W. Schafer, *Digital Processing of Speech Signals*, New Jersey 1978, 476–489.

2. Trawińska A. (2009): Analiza mowy i nagrań, (w:) Postępy w naukach sądowych, Kała M. (red.), Wydawnictwo Instytutu Ekspertyz Sądowych, Kraków, 117–134.

3. K. Klus, A. Trawińska, *Forensic Speaker Identification by the Linguistic-Acoustic Method in KEÚ AND IES* (w:) Problems of Forensic Sciences 2009, vol. LXXVIII, 160–174.

4. R. Tadeusiewicz, *Sygnał Mowy*, Wydawnictwo Komunikacji I Łączności, Warszawa 1988, 161–172.

5. H. Sakoe, S. Chiba, *Dynamic Programming Algorithm Optimization for Spoken Word Recognition*, (w:) IEEE Transatctions of Acoustics, Speech and Signal Processing, vol. ASSP-26, No. 1, Luty 1978.

6. B. Ziółko, M. Ziółko, *Przetwarzanie mowy,* Wydawnictwa AGH, 2011.

7. S.B. Davis and P. Mermelstein, *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, IEEE Transactions on Acoustics, Speech and Signal Processing, 1980, vol. ASSP-28, pp. 357–366, no. 4.

8. J. Makhoul, *Spectral linear prediction: properties and applications*, IEEE Transcations, 1975, vol. ASSP-23, pp. 283–296.

9. B. Ziółko, W. Kozłowski, M. Ziółko, R. Samborski, D. Sierra, J. Gałka, *Hybrid Wavelet-Fourier-HMM Speaker Recognition*, *International Journal of Hybrid Information Technology,* vol. 5, No. 4, October, 2011.

10. K. Malik, *Uwarunkowania językowe i pozajęzykowe procesu rozpoznania mówcy przez świadka ze słyszenia***,** Problemy Kryminalistyki nr 271/2011.

11. J. Rzeszotarski, Identyfikacja mówcy celowo zniekształcającego wypowiedzi, Problemy Kryminalistyki nr 255/2007.

12. J. Aach, G. M. Church, Aligning gene expression time series with time warping algorithms, Bioinformatics (2001) 17 (6), 495–508.

13. P. Żelasko, Półautomatyczne rozpoznawanie mówców w kryminologii, Praca inżynierska wykonana w Katedrze Elektroniki, WIEiT AGH, Kraków 2013.

Streszczenie

W artykule omówiono problemy związane z identyfikacją mówcy i przedstawiono propozycję procedury ułatwiającej proces identyfikacji w części akustycznej. Koncepcja opiera się na metodach programowania dynamicznego, a w szczególności algorytmu znanego jako DTW (dynamic time warping). Przeprowadzone zostały testy wskazujące na przydatność proponowanej procedury przy próbie ustalenia, które samogłoski oraz formanty pozwalają dostatecznie zróżnicować mówców, dostatecznie indywidualizując każdego z nich.

Słowa kluczowe: *analiza nagrań, DTW, formant, identyfikacja mówcy, metoda językowo-pomiarowa*

Summary

The paper discusses issues concerning forensic speaker identification and proposes a procedure simplifying the process of speaker identification in the acoustic scope. The idea revolves around methods of dynamic programming, especially the DTW (dynamic time warping) algorithm described further in the paper. Tests which were carried out demonstrated the usefulness of the suggested procedure when trying to determine which vowels and formants are the best differentiating and characterizing a speaker.

Keywords: *recording examination, DTW, formant, speaker identification, linguistic-acoustic method*