Robert Wielgat[1]*, Anita Lorenc[2]

[1] Państwowa Wyższa Szkoła Zawodowa w Tarnowie, Instytut Politechniczny, Zakład Elektroniki i Telekomunikacji
State Higher Vocational School in Tarnow, Polytechnic Institute, Electronics and Telecommunictions
[2] Uniwersytet Marii Curie-Skłodowskiej w Lublinie, Zakład Logopedii i Językoznawstwa Stosowanego
Uniwersytet Marii Curie-Skłodowskiej in Lublin, Department of Speech Therapy and Applied Linguistics
*corresponding author: rwielgat@poczta.onet.pl

Inwersja mowy za pomocą nieliniowej transfomacji czasowej
Inversion of speech by non-linear transformation of temporary

Streszczenie

Artykulografia Elektromagnetyczna (ang. Electromagnetic Articulography – EMA) jest precyzyjną metodą diagnozy narządów mowy dokonywaną za pomocą czujników pola elektromagnetycznego umieszczonych głównie na języku. Pomimo swej precyzji badanie jest dość uciążliwe dla mówcy, dlatego poszukuje się różnych innych metod diagnozy. Jedną z nich jest inwersja mowy polegająca na estymacji ruchów języka na podstawie nagrań dźwiękowych. W niniejszym artykule opisano wstępne badania nad inwersją mowy z wykorzystaniem nieliniowej transformacji czasowej (ang. DTW). Jako metodę parametryzacji sygnału mowy wybrano współczynniki mel-cepstralne (ang. MFCC). Obliczono i przedyskutowano błędy estymacji na przykładzie słów języka polskiego.

Słowa kluczowe: inwersja mowy; nieliniowa transformacja czasowa; parametry mel-cepstralne

Abstract

Electromagnetic Articulography (EMA) is a precise method for speech articulators assessment which is carried out by sensors placed mainly on the tongue. Various methods are being developed in order to avoid the assessment by EMA sensors. One of them is speech inversion. Here preliminary research on speech inversion based on dynamic time warping (DTW) method has been described. Mel-frequency cepstral coefficients (MFCC) method has been chosen as the acoustic speech signal parametrization method. Root mean square errors (RMSE) of the evaluation have been presented and discussed.

Keywords: inversion of speech; non-linear transformation of the time; mel-cepstral parameters

Wstęp

Artykulografia elektromagnetyczna jest metodą badania logopedycznego stosowaną na świecie od ponad dwóch dekad. Służy ona do diagnozowania narządów mowy, w szczególności do śledzenia ruchu języka podczas wymawiania badanych fonemów [1, 2]. Metoda ta jest bardzo precyzyjna i efektywna jakkolwiek niezbyt przyjemna dla osoby badanej. Z tego powodu są podejmowane próby nieinwazyjnego badania ruchów języka. Jedną z nich jest inwersja mowy polegająca na estymacji trajektorii ruchu języka na podstawie sygnału akustycznego oraz obrazów wideo [3, 4]. Najpowszechniej stosowanymi metodami do inwersji mowy wydają się być metody oparte na ukrytych modelach Markowa (ang. HMM) [5, 6].

W artykule zaprezentowano inne podejście do inwersji mowy. Opisywana tutaj metoda jest oparta na paradygmacie programowania dynamicznego. W celu wykorzystania metody należy nagrać zestaw wzorców. Wzorce są sekwencjami obserwacji (wektorami cech) otrzymanymi z dźwiękowego sygnału mowy. Każdej obserwacji towarzyszą położenia sensorów EMA umieszczonymi na języku i wargach będące próbkami sygnałów artykulograficznych. W celu otrzymania estymowanych położeń sensorów dla nagranego dźwiękowo słowa należy wyznaczyć dopasowanie czasowe między tym słowem a najbardziej podobnym wzorcem tej samej klasy, co nagrane słowo. Następnie położenia sensorów skojarzone z obserwacjami (wektorami cech) wzorca są przypisywane dopasowanym czasowo obserwacjom nagranego słowa. Jako metodę parametryzacji wybrano metodę MFCC [7].

Materiał i metody

W celu dokonania inwersji mowy należy nagrać zbiór sygnałów audio i towarzyszących im sygnałów artykulograficznych. Sygnały audio reprezentują słowa, które są konwertowane na wzorce (sekwencje obserwacji) za pomocą metody MFCC. Sygnały artykulograficzne są sygnałami cyfrowymi nagrywanymi za pomocą artykulografu elektromagnetycznego (EMA). Poszczególne próbki sygnałów artykulograficznych stanowią położenia sensorów przypisane do odpowiadających im obserwacji. Kompletny opis systemu akwizycji do nagrywania sygnałów audio oraz rejestracji sygnałów artykulograficznych opisano w pozycji [8]. Zbiór wzorców wraz

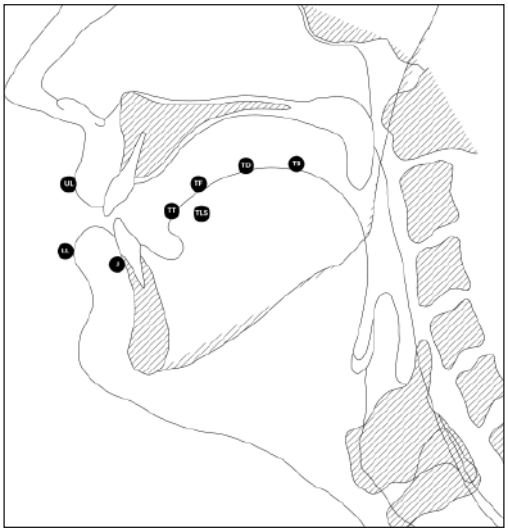
z zarejestrowanymi położeniami czujników z artykulografu jest dzielony na zbiór uczący i testowy. Położenia sensorów dla słów ze zbioru testowego są estymowane za pomocą metody DTW. Jakość estymacji jest oceniana na podstawie błędu rms (ang. root mean square error – RMSE) między pomierzonymi a estymowanymi położeniami sensorów.

Artykulografia elektromagnetyczna

Badania przeprowadzono za pomocą artykulografu elektromagnetycznego AG 500. Głównymi elementami EMA jest sześć dużych cewek wytwarzających zmienne pole elektromagnetyczne o trzech składowych częstotliwościowych. Pole elektromagnetyczne indukuje zmienne prądy w małych cewkach czujników przyklejonych do artykulatorów mowy, głównie do języka oraz warg. Indukowane prądy są analizowane za pomocą szybkiej transformaty fouriera (FFT). Na podstawie analizy jest obliczana energia każdej składowej częstotliwościowej prądu zaindukowanego w czujnikach. Na podstawie wartości energii jest obliczane położenie oraz orientacja czujników w przestrzeni 3D.

Do wyznaczenia położenia ruchomych artykulatorów podczas wypowiadania słów języka polskiego należało przykleić 12 czujników EMA do wybranych uprzednio punktów rozmieszczonych w różnych punktach twarzy i ruchomych artukulatorów. Czujniki przyklejano za pomocą nietoksycznego kleju. Rozmieszczenie czujników na wargach i języku stanowiące ruchome artykulatory mówcy zostało pokazane na Rysunku 1.

Oprócz sensorów rozmieszczonych na wargach i języku konieczne było umieszczenie na głowie mówcy trzech sensorów odniesienia pozwalających na korektę niepożądanych ruchów głowy w trakcie badania. Sensory zostały umieszone w punktach związanych z kościami czaszki. Dwa z nich zostały umieszczone na wyrostkach sutkowatych za uszami a jeden w zagłębieniu u nasady nosa. Takie rozmieszczenie sensorów odniesienia redukowało do minimum ich przemieszczanie się podczas badania. Dwa sensory kontrolujące ruchy warg (LL, UL) zostały umieszczone w płaszczyźnie symetrii twarzy. Kolejne cztery czujniki zostały rozmieszczone wzdłuż linii środkowej języka: jeden czujnik na czubku języka (TT), jeden w obszarze postdorsalnym z tyłu języka (TB) i dwa następne (TF, TD), które rozłożono równomiernie między skrajnymi czujnikami (TT, TB). Ostatni czujnik na języku (TLS) zo-



Rysunek 1. Rozmieszczenie czujników
na ruchomych artykulatorach mówcy

stał przyklejony z lewej strony z wierzchu języka pomiędzy czujnikami TT oraz TF. Rozmieszczenie czujników na języku mówcy pokazuje Rysunek 2.

Ostatni czujnik (J) został naklejony na granicy między dziąsłami i dolnymi siekaczami w celu kontrolowania ruchów rzuchwy (Rysunek 1). Jeden z czujników służył do wykonania obrysu podniebienia twardego i częściowo miękkiego.

Rysunek 2. Sensor placement in the speaker MK (female)

Parametryzacja sygnału mowy

Akustyczny sygnał mowy nagrano z częstotliwością próbkowania 96 kHz. Jako cechy z sygnału mowy ekstrahowano parametry mel-cepstralne (MFCC). Sposób obliczania parmetrów MFCC jest oparty na modelowaniu przetwarzania sygnału akustycznego podobnym do zachodzącego w ślimaku narządu słuchu. Zgodnie z tym sposobem parametry MFCC były obliczane w eksperymentach w następujących krokach:

1) blokowanie sygnału w ramki i okienkowanie oknem Hamminga;
2) wyznaczanie szybkiej transformaty Fouriera ze zokienkowanych ramek;
3) obliczanie mocy FFT w równomiernie rozmieszczonych w skali melowej pokrywających się na długości 50% pasmach, w których wartości mocy prążków FFT są ważone za pomocą funkcji trójkątnych. Liczba pasm jest parametrem algorytmu. Konwersja z liniowej skali częstotliwości na skalę melową i vice versa są dane za pomocą równań:

$$f_{mel} = 2595 \log_{10}\left(1 + f_{Hz} / 700\right) \tag{1}$$

$$f_{Hz} = 700 \cdot \left(10^{f_{mel}/2595} - 1\right) \tag{2}$$

4) Obliczenie logarytmu mocy (obliczonej w p. 4) w pasmach melowych.

5) Przeprowadzenie DCT na wektorach logarytmów mocy z p. 5. (n = 0, 1,…, q-1):

$$X(n) = c(n) \sum_{k=0}^{K-1} \ln(S_k) \cos\left(\frac{\pi(2k+1)n}{2K}\right) \tag{3}$$

$$c(0) = \sqrt{\frac{1}{K}} \quad , \quad c(n) = \sqrt{\frac{2}{K}} \qquad \text{for n = 1,…, q-1}$$

gdzie: S_k – logarytm widmowego współczynnika mocy w k-tym paśmie częstotliwościowym; K – liczba pasm częstotliwościowych; q – liczba współczynników MFCC.

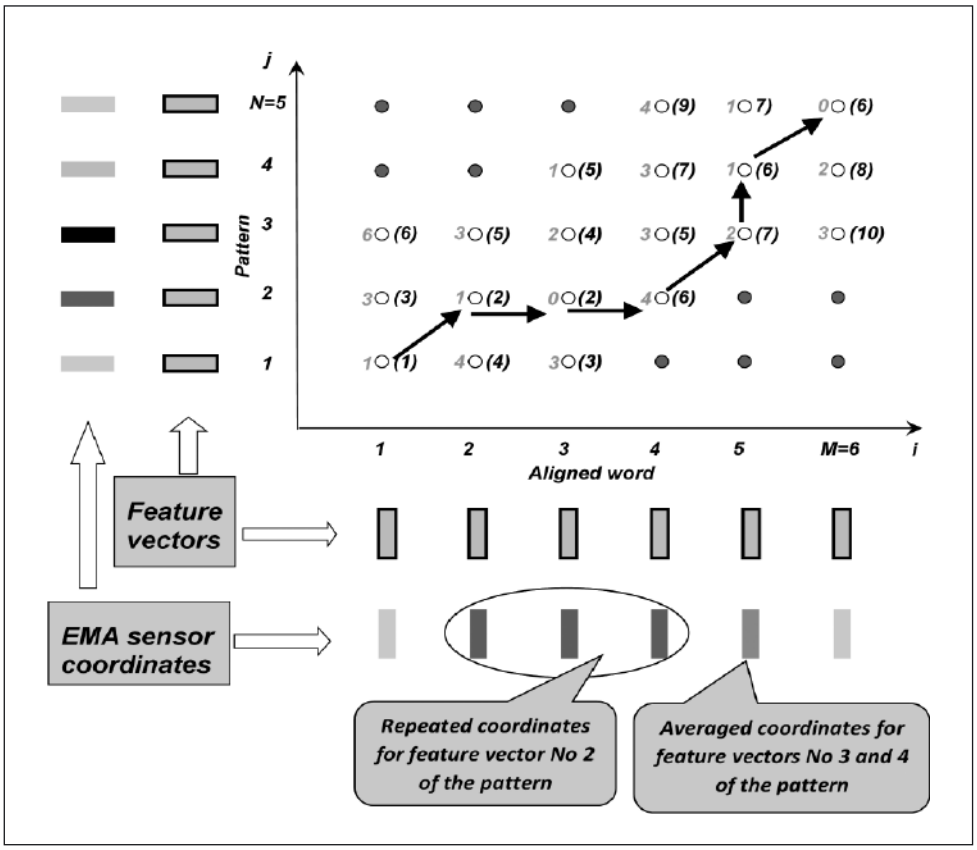
Inwersja mowy

Metodą inwersji mowy zastosowaną w badaniach jest nieliniowa transformacja czasowa. Metoda jest oparta na paradygmacie programowania dynamicznego. W celu użycia metody powinien zostać nagrany zbiór wzorców. Wzorce są sekwencjami obserwacji (wektory cech MFCC) ekstrahowanymi z sygnału mowy. Każdej obserwacji towarzyszą nagrywane jednocześnie położenia sensorów umieszczonych na języku i wargach. Badane słowo jest porównywane z wzorcem tej samej klasy poprzez wyznaczenie optymalnego dopasowania czasowego między nimi. Optymalne dopasowanie jest obliczane za pomocą algorytmu DTW [9, 10]. Aby wyznaczyć optymalne dopasowanie czasowe za pomocą metody DTW należy najpierw obliczyć macierz lokalnych odległości euklidesowych pomiędzy wszystkimi wektorami cech (MFCC) wzorca oraz słowa badanego. Następnie dla każdego punktu macierzy odległości lokalnych jest obliczana odległość zakumulowana zgodnie ze wzorem (4):

$$d_A(i,j) = D(i,j) + \cdots$$
$$\cdots + min\{d_A(i,j-1), d_A(i-1,j-1), d_A(i-1,j)\} \quad (4)$$

W każdym punkcie i, j są zapisywane współrzędne (indeksy) poprzedniej minimalnej odległości d_A w celu śledzenia wstecznego optymalnej ścieżki. Na końcu wybierana jest minimalna odległość zakumulowana w obszarze górnego prawego rogu. Ta odległość zakumulowana po normalizacji przez podzielenie przez długość przekątnej macierzy odległości lokalnych stanowi odległość DTW. Wzorzec z najmniejszą odległością DTW jest wybierany do dalszego przetwarzania.

Po wyborze najbardziej podobnego wzorca do dalszego procedowania wyznaczana jest optymalna ścieżka. Optymalna ścieżka określa również najlepsze dopasowanie czasowe między wektorami cech najlepszego wzorca a wektorami cech słowa, dla którego mają być estymowane położenia czujników. Położenie sensora skojarzone z wektorem cech badanego słowa jest uznawane za takie samo jak położenie sensora skojarzonego z dopasowanym czasowo wektorem cech wzorca. Gdy kilka wektorów cech wzorca jest dopasowanych czasowo z wektorem cech badanego słowa, wówczas aby uzyskać położenie sensorów dla wektora cech z badanego słowa, jest obliczane średnie położenie sensorów skojarzonych z dopasowanymi czasowo wektorami cech wzorca. Proces przypisywania położeń sensorów do poszczególnych



Rysunek 3. Metoda DTW używana do inwersji mowy

wektorów cech został pokazany na Rysunku 3. Po obliczeniu położenia sensorów zostały wyznaczone minimalne oraz maksymalne bezwzględne błędy estymacji oraz błędy rms między rzeczywistym, a estymowanym położeniem sensorów (Tabela 1.).

Wyniki

Jako materiał badawczy nagrano za pomocą 16-kanałowego rejestratora audio sygnał mowy z relatywnie wysoką częstotliwością próbkowania 96 kHz. Do badań opisywanych w artykule wykorzystano nagrania tylko z jednego kanału. Na podstawie sygnału mowy wyznaczono wektory współczynników MFCC. Po akwizycji sygnału, był on dzielony na ramki o czasie trwania 21,33 ms i czasie zachodzenia ramek na siebie 16,33 ms. Wybór takiej długości ramki był podyktowany uzyskaniem liczby próbek w ramce będącej potęgą liczby 2 w celu dokonania FFT. Natomiast czas

zachodzenia ramek na siebie był podyktowany uzyskaniem tej samej częstotliwości ramki co częstotliwość próbkowania sygnałów z sensorów EMA, która wynosi 200 Hz. Ramki sygnału okienkowano oknem Hamminga w celu zminimalizowania efektu tzw. „przecieku widma". Liczba współczynników MFCC wynosiła 12. Liczba pasm częstotliwościowych wynosiła 34. Prążki częstotliwościowe w pasmach były ważone za pomocą filtrów trójkątnych. Długość transformaty FFT wynosiła 2048.

Tabela 1.
Błędy inwersji mowy

Czujnik	Kierunek ruchu	Błąd [mm] (*objaśnienia poniżej tabeli*)		
		min	średni	maks
UL	X	0,00	0,12	0,31
	Y	0,01	0,16	0,36
	Z	0,01	0,95	1,44
LL	X	0,00	0,00	0,00
	Y	0,01	0,78	2,31
	Z	0,00	0,28	0,81
JAW	X	0,01	0,68	1,89
	Y	0,00	1,29	3,42
	Z	0,00	0,00	0,00
TB	X	0,00	2,90	11,24
	Y	0,02	1,82	3,17
	Z	0,01	1,97	3,43
TT	X	0,00	0,92	2,04
	Y	0,00	0,71	2,01
	Z	0,01	2,72	7,07
TD	X	0,00	1,06	2,36
	Y	0,00	0,00	0,00
	Z	0,01	2,43	4,57

TF	X	0,03	2,50	6,11
	Y	0,0	6,33	13,16
	Z	0,00	2,66	11,04
TL	X	0,00	0,00	0,00
	Y	0,00	2,56	9,84
	Z	12,33	290,43	344,38

min – minimalny błąd bezwzględny
średni –błąd rms
maks – maksymalny błąd bezwzględny

Na głowie i języku zostało umieszczonych 12 czujników EMA tak jak to opisano w rozdziale II.A. Do eksperymentu wybrano nagrania słowa „Andrzej". Jedno nagranie wybrano jako wzorcowe, a drugie jako testowe. Dokonano inwersji mowy na podstawie słowa wzorcowego za pomocą metody DTW. Rezultaty zaprezentowano w Tabeli 1. Wyniki pokazują, że dla niektórych sensorów oraz kierunków ich przemieszczania się można osiągnąć minimalne błędy na poziomie wartości 0. Z drugiej strony dla niektórych sensorów i kierunków ich przemieszczania się zaobserwowano duże błędy rms oraz maksymalne błędy bezwzględne o wartościach powyżej 5 mm. Błędy te nie wynikają z naturalnych, fizjologicznych różnic w wymowie tych samych fonemów przez tego samego mówcę [11], mało jest również prawdopodobne aby wynikały one z niedoskonałości prezentowanej metody inwersji mowy. Najbardziej prawdopodobnym źródłem wspomnianych błędów wydają się dosyć często obserwowane zakłócenia podczas akwizycji sygnałów z czujników EMA. Dlatego przypadki, gdzie błąd rms jest większy niż 5 mm nie zostały uwzględnione w ocenie dokładności prezentowanej metody inwersji mowy. Pogłębioną analizę przyczyn wspomnianych błędów planuje się przeprowadzić w ramach przyszłych badań.

Wnioski

W wyniku badań została wstępnie sprawdzona metoda inwersji mowy oparta na nieliniowej transformacji czasowej. Wyniki pokazały, że średni bezwzględny błąd estymacji był mniejszy niż 3,1 mm. Natomiast dla niektórych czujników oraz kierunków ich poruszania się osiągnięto prawie perfekcyjną estymację z błędem bliskim 0.

Jednocześnie stwierdzono, że zakłócenia podczas akwizycji sygnałów z sensorów EMA mają duży wpływ na jakość inwersji mowy. Eliminacja tych zakłóceń będzie jednym z ważniejszych zadań w przyszłych badaniach.

Dalsze badania obejmują również bardziej wszechstronne testowanie inwersji mowy za pomocą metody DTW. Planowane jest również użycie innych metod min. opartych na ukrytych modelach Markowa (HMM) oraz sieciach Bayesa.

Podziękowania

Opisane w artykule badania wykonano w ramach grantu Nr 2012/05/E/HS2/03770 zatytułowanego. Współczesna wymowa polska. Badanie z wykorzystaniem trójwymiarowej artykulografii elektromagnetycznej". Kierownikiem projektu jest Anita Lorenc. Projekt sfinansowano ze środków Narodowego Centrum Nauki w oparciu o decyzję Nr DEC-2012/05/E/HS2/03770.

Piśmiennictwo

[1] Perkell J.S., Cohen M.H., Svirsky M.A., Matthies M.L., Garabieta I., Jackson M. T. Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements. JASA, 1992, 92(6), 3078-96.

[2] Król D., Lorenc A., Święciński R. Detecting Laterality and Nasality in Speech with the Use of a Multi-Channel Recorder. Proceedings of the 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, 5147-51.

[3] Beskow J., Engwall O., Granström B. Simultaneous measurements of facial and intraoral articulation. Proceedings of Fonetik 2003. Dept. of Linguistics, Stockholm University, 2003, 57-60.

[4] Kjellström H., Engwall O. Audiovisual to articulatory inversion. Speech Communication, 2009, 51(3), 195-209.

[5] Richmond K. Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion. Advances in Nonlinear Speech Processing, Lecture Notes in Computer Science 2007, 4885, 263-72.

[6] Hueber T., Ben Youssef A., Bailly G., Badin P., Eliséi F. Cross-speaker Acoustic-to-Articulatory Inversion using Phone-based Trajectory HMM for Pronunciation Training. Proceedings of Interspeech, Portland, USA, 2012.

[7] Makowski R., Świętojański P., Wielgat R. Automatyczne rozpoznawanie mowy. Chapter 14 In book: Cyfrowe Przetwarzanie Sygnałów w Telekomunikacji. Podstawy, multimedia, transmisja. Publisher: Wydawnictwo Naukowe PWN - Red: Zielinski, T., Korohoda, P., Rumian, R. 2014, 522-30.

[8] Mik Ł., Wielgat R., Lorenc A., Król D., Święciński R., Jędryka R. Multimodal Speech Data Acquisition with the Use of EMA Fast-speed Video Cameras and a Dedicated Microphone Array. 23rd International Conference Mixed Design of Integrated Circuits and Systems (MIXDES), Łódź, Poland, June 2016.

[9] Rabiner L.R., Rosenberg A., Levinson S. Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition", IEEE Trans. Acoust., Speech, Signal Processing, 1978, 26, 575-82.

[10] Kuhn M.H., Tomaschewski H.H. Improvements in Isolated Word Recognition. IEEE Trans. Acoust., Speech, Signal Processing, 1983, 31(1), 157-67.

[11] Lorenc A. Wymowa normatywna polskich samogłosek nosowych i spółgłoski bocznej, (rozdział 4.4). Dom wydawniczy ELIPSA, Warszawa 2016.