Projekt koncepcyjny bazy danych do przechowywania nagrań z badań artykulograficznych mowy polskiej

Conceptual design of a database to store recordings from articulographic studies of Polish speech

Robert Wielgat[a,*], Rafał Jędryka[a], Anita Lorenc[b,c], Łukasz Mik[a], Daniel Król[a]

[a] *State Higher Vocational School in Tarnów, Mickiewicza 8, 33-100 Tarnów, Poland*
[b] *Maria Curie-Skłodowska University, Department of Speech Therapy and Applied Linguistics, Sowińskiego 17, 20-040 Lublin, Poland*
[c] *Warsaw University, Institute of Applied Polish Studies, Department of Speech and Language Therapy and Voice Production, Krakowskie Przedmieście 26/28, 00-927 Warszawa, Poland*

Article history:
Received 1 December 2017
Received in revised form
22 December 2017
Accepted 23 December 2017
Available online 27 December 2017

Streszczenie

W artykule opisano strukturę i funkcjonalność bazy danych artykulograficznych do przechowywania danych z badań przeprowadzanych z wykorzystaniem artykulografu elektromagnetycznego, kamery akustycznej i 3 kamer wideo. Baza danych umożliwia selektywne pobieranie różnych typów danych, w szczególności dotyczących mówcy, sesji nagraniowej, nagrań oraz eksperymentów. Opisano strukturę i budowę bazy danych. Przedstawiono również potencjalne przyszłe zastosowania do przeprowadzania analiz statystycznych oraz w eksperymentach dotyczących inwersji mowy z wykorzystaniem modeli sieci Bayesa.

Słowa kluczowe: artykulografia elektromagnetyczna, bazy danych, sieci Bayesa, inwersja mowy, kamera akustyczna, fonetyka artykulacyjna, fonetyka akustyczna

1. Wstęp

Relacyjne bazy danych znalazły zastosowanie w prawie wszystkich dziedzinach życia w ciągu ostatnich trzech dekad. Specjalną klasą baz danych są bazy danych wykorzystywane do badań naukowych. W niniejszej pracy opisano koncepcję naukowej bazy nagrań wykonanych za pomocą elektromagnetycznego artykulografu (EMA) i urządzeń towarzyszących.

Badania za pomocą artykulacji elektromagnetycznej mają swoją ponad dwudziestoletnią historię [1]. Służą one do oceny aktywności ruchowej narządów mowy, w szczególności ruchu języka. Nagrania dokonuje się poprzez śledzenie trajektorii ruchu małych czujników umieszczonych w różnych punktach narządów mowy. Nagraniom artykulatorów zwykle towarzyszą nagrania uzyskane z innych urządzeń, np. kamer wideo [2, 3], rejestratorów audio [4, 5] lub kamery akustycznej [6]. Do tej pory na świecie powstało kilka baz danych artykulacyjnych. Historycznie najstarszą jest baza danych MOCHA-TIMIT [7], która oprócz nagrań z EMA przechowuje również dane uzyskane z rejestratora dźwięku, kamery wideo, laryngografu oraz elektropalatografu (EPG). Baza danych zawiera nagrania 460 różnych zdań w języku angielskim wypowiadanych przez 1 mężczyznę i 1 kobietę. Inna baza danych USC-TIMIT [8] zawiera dane uzyskane z EMA, nagrania dźwiękowe i obrazy rezonansu magnetycznego w czasie rzeczywistym dla tych samych 460 zdań, jak w bazie danych MOCHA-TIMIT. Zdania nagrane przez EMA zostały wypowiedziane przez 2 mężczyzn i 2 kobiety posługujących się amerykańską odmianą języka angielskiego. Wspomniane bazy nagrań różnią się sprzętem użytym do nagrań. W przypadku bazy USC-TIMIT nie ma nagrań dokonywanych za pomocą kamery wideo, laryngografu oraz elektropalatografu. Są za to obrazy rezonansu magnetycznego. Różne są również typy użytych artykulografów. W przypadku bazy USC-TIMIT jest to artykulograf Northern Digital natomiast w przypadku bazy MOCHA-TIMIT jest to artyklograf Carstens Medizinelektronik. Jeszcze jedną bazą danych opartą na zdaniach z bazy MOCHA-TIMIT jest baza TORGO zawierająca nagrania osób cierpiących na dyzartię [9]. Słownik wypowiadanych słów zawiera sylaby i samogłoski, około 800 izolowanych słów, ponad 600 zdań czytanych z monitora wliczając w to zdania z bazy TIMIT, zdania wypowiadane w mowie spontanicznej. Baza TORGO obejmuje nagrania EMA, audio i wideo dla wypowiedzi 4 mężczyzn i 3 kobiet. Sprzęt użyty do nagrań obejmował artykulograf AG 500, dwie kamery wideo, mikrofon nagłowny oraz 8-mikrofonową macierz typu Acoustic Magic Voice Tracker. Baza danych zawierająca znacz-

*Corresponding author: rwielgat@poczta.onet.pl

nie więcej zdań niż wyżej wymienione bazy danych to zbiór danych mngu0 [4] zawierający nagrania około 2100 zdań. Jednak dane pochodzą tylko od jednego mówcy posługującego się językiem angielskim. Nagraniom z artykulografu towarzyszą nagrania wideo oraz nagrania audio dokonywane za pomocą pojemnościowego mikrofonu hiperkardioidalnego oraz mikrofonu optycznego. Oprócz tych nagrań w bazie znajdują się nagrania inwentarza głosek mówcy otrzymane za pomocą rezonansu magnetycznego oraz zeskanowane przestrzennie odlewy dentystyczne dolnej i górnej szczęki. Bazą, która obejmuje obecnie najwięcej mówców jest baza EMA-MAE [10]. Znalazły się w niej nagrania 20 mówców obydwu płci wypowiadających słowa w amerykańskiej odmianie języka angielskiego oraz 20 mówców obydwu płci wypowiadających słowa w języku angielskim z akcentem mandaryńskim. Każdy z mówców wypowiadał 330 izolowanych słów, wybrane zdania z bazy TIMIT oraz połączone zdania z jednego akapitu. Nagraniom z artykulografu towarzyszyły nagrania dźwiękowe rejestrowane za pomocą mikrofonu pojemnościowego ustawionego w odległości 1 m od mówiącego. Oprócz baz nagrań artykulograficznych dla języka angielskiego zostały również stworzone bazy nagrań dla innych języków. Jedną z nich jest baza Qualisys-Movetrack obejmująca nagrania dla języka szwedzkiego [11]. Autorzy bazy dokonali nagrań jednej kobiety za pomocą EMA, magnetofonu kasetowego DAT oraz systemu przechwytywania ruchu twarzy (ang. motion capture) MacReflex szwedzkiej firmy Qualisys. Baza nagrań obejmowała 270 zdań oraz około 180 wyrazów wypowiedzianych w języku szwedzkim. Inną bazą nagrań jest baza nagrań dla języka estońskiego [12] obejmująca nagrania jednego mężczyzny i jednej kobiety za pomocą artykulografu oraz EMA. Baza ta obejmuje również dokonywane osobno nagrania za pomocą EGG, EPG oraz rejestratora audio.

Dla języka polskiego pionierskie prace w zakresie nagrań EMA przeprowadzono w Berlin Zentrum für Allgemeine Sprachwissenschaft. Były to nagrania obejmujące rejestrację sygnału EMA oraz nagrania audio dla splastycznych artykulacji wariacyjnych spółgłosek wargowych /pj/ i /bj/ [13, 14]. Od 2009 r. nagrania artykulograficzne dla języka polskiego są prowadzone przy użyciu artykulografu elektromagnetycznego AG500 firmy Carstens Medizinelektronik w Zakładzie Logopedii i Językoznawstwa Stosowanego UMCS w Lublinie. Nagrane zbiory danych artykulograficznych obejmują nagrania dwóch osób z wymową normatywną i wadliwą [15, 16] oraz nagrania do testów porównawczych opisujące podstawy artykulacji podczas mówienia po polsku i angielsku u mówców, dla których polski jest językiem ojczystym [17]. Nagraniom EMA towarzyszą nagrania audio oraz nagrania z kamery wideo.

Opisywana w niniejszej publikacji baza POLEMAD (akronim od angielskiej nazwy „**POL**ish **E**lectro**M**agnetic **A**rticulography **D**atabase") powstała w rezultacie kontynuacji opisanych powyżej nagrań prowadzonych w UMCS w Lublinie. W odróżnieniu od innych baz nagrań artykulograficznych baza POLEMAD zawiera nagrania z 3 szybkich kamer wideo, dzięki czemu jest możliwe otrzymywanie trójwymiarowego obrazu twarzy oraz zawiera nagrania dźwiękowe z 16-kanałowego rejestratora audio, co pozwala na generowanie obrazów twarzy z mapą natężenia dźwięku. Ponadto jest największą bazą nagrań artykulograficznych dla języka polskiego i jedną z największych baz na świecie pod względem liczby nagranych mówców, ponieważ zawiera nagrania pochodzące od 10 mężczyzn oraz 10 kobiet. Baza POLEMAD jest również prawdopodobnie jedną z niewielu, która posiada relacyjną strukturę danych. W literaturze światowej brak jest informacji na ten temat.

2. Opis bazy danych

W bazie POLEMAD będą przechowywane dane dotyczące nagrań artykulograficznych pochodzące od 20 mówców: 10 kobiet i 10 mężczyzn z normatywnym typem wymowy polskiej. Każdy z mówców wypowiedział około 400 izolowanych słów z listy wyrazowej. Lista wyrazowa zawiera izolowane słowa języka polskiego zawierające wszystkie głoski języka polskiego w śródgłosie w pozycji akcentowanej. Wypowiedzi mówców nagrywano za pomocą specjalnego systemu akwizycji danych artykulacyjnych [18]. Nagrania dokonywane za pomocą systemu obejmują zarejestrowane położenia i kąty nachyleń czujników artykulografu elektromagnetycznego AG 500, nagrania video z 3 szybkich kamer Gazelle GZL-CL-22C5M-C firmy Point Grey oraz 16-kanałowe nagrania dźwiękowe dokonywane za pomocą rejestratora audio z kołową macierzą mikrofonową. Szczegółowy opis danych przechowywanych w bazie zawarto w rozdziale 2.1.

Baza danych będzie posiadać relacyjną strukturę tabel i ma ułatwiać selektywne pobieranie przechowywanych w niej informacji. Do implementacji zostanie wykorzystany system zarządzania bazą danych PostgreSQL. Struktura bazy została przestawiona w rozdziale 2.2. Informacje są pobierane z bazy przede wszystkim w celu wizualnej i odsłuchowej analizy zaobserwowanych zależności fonetycznych i fonologicznych w poszczególnych głoskach języka polskiego. Selektywne pobieranie danych z bazy pozwala również na łatwe wykonywanie różnego rodzaju opracowań naukowych obejmujących analizy statystyczne oraz doświadczenia nad inwersją mowy. Szczególnym rodzajem inwersji mowy są metody wykorzystujące sieci Bayesa. Wykorzystanie bazy danych do eksperymentów naukowych opisano w rozdziale 3.

2.1. Rodzaje danych przechowywanych w bazie

Projekt bazy POLEMAD zakłada przechowywanie różnych typów danych: od danych typu tekstowego po pliki dźwiękowe, obrazy oraz filmy.

W dalszej części rozdziału zostanie przedstawiony bardziej szczegółowy opis wybranych rodzajów danych.

2.1.1. Dane mówcy

O mówcy będą przechowywane podstawowe informacje, które nie są klasyfikowane jako dane osobowe. W zakres danych wchodzą takie informacje jak: inicjały, płeć, rok urodzenia mówcy, wynik badania logopedycznego oraz wynik badania audiometrycznego. Dane anatomiczne oraz wyniki badań: logopedycznego i audiometrycznego pozwalają na ocenę, czy dany mówca posiada wymowę normatywną.

2.1.2. Dane z artykulografu elektromagnetycznego

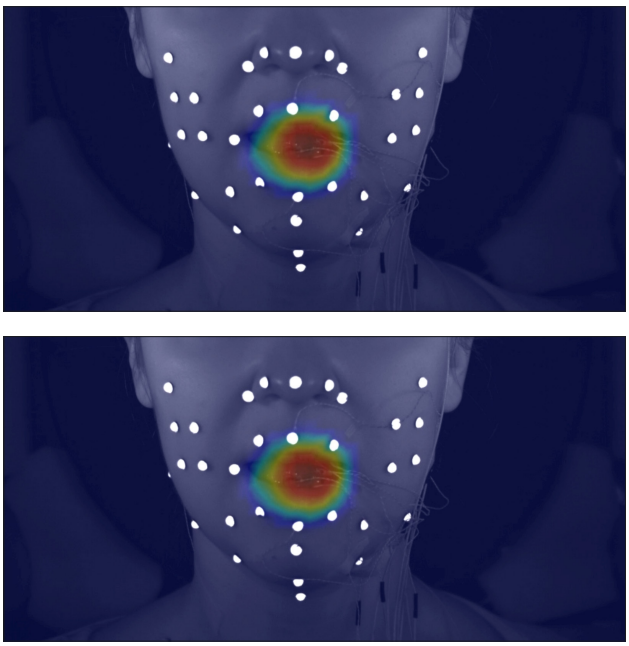
Dane rejestrowane z artykulografu elektromagnetycznego stanowiły zarejestrowane co 5 ms położenia oraz kąty nachylenia 12 sensorów artykulografu. Położenia sensorów były rejestrowane za pomocą artykulografu AG 500 firmy Carstens Medizinelektronik. Położenia sensorów były zapisywane jako współrzędne x, y, z w układzie kartezjańskim związanym z kabiną artykulografu. Początek układu współrzędnych pokrywa się ze środkiem kabiny. Rejestrowano 2 kąty nachylenia φ oraz θ. Po zarejestrowaniu położeń czujników oraz kątów nachylenia dokonywano korekty ruchów głowy, czyli transformacji położeń czujników oraz kątów nachylenia z układu współrzędnych związanym z kabiną artykulografu na układ współrzędnych związany z czaszką osoby mówiącej. Układ współrzędnych związany z czaszką mówiącego wyznaczały trzy czujniki referencyjne umieszczone we względnie nieruchomych względem czaszki punktach, czyli u nasady nosa oraz na wyrostkach sutkowatych za uszami. Transformacji dokonywano za pomocą metody wykorzystującej algorytm Newtona-Raphsona [19, 20].

2.1.3. Nagrania wideo

Nagrania wideo otrzymano z systemu wizyjnego złożonego z trzech kamer: frontalnej, prawej i lewej (Rys. 1). Do nagrań wideo użyto kamer firmy Point Grey – model Gazelle GZL-CL-22C5M-C. Każda z tych kamer jest wyzwalana z częstotliwością 200 Hz (częstotliwość próbkowania AG500) w celu pełnej synchronizacji z artykulografem. Podczas sesji nagraniowych dane wideo są zapisywane w postaci nieskompresowanej, w formacie AVI. Docelowo w bazie są umieszczone nagrania z kompresją Motion JPEG, co pozwoliło zachować współczynnik fps na poziomie 200 bez utraty rozdzielczości, która aktualnie wynosi 1024 × 1024 piksele. Uzyskano w ten sposób zmniejszenie rozmiaru danych wideo do poziomu ok. 4% w stosunku do stanu początkowego. Jakość nagrań po kompresji jest dobra i nie wpływa negatywnie na realizację algorytmów przetwarzania obrazów, służących do wyszukiwania i śledzenia ruchu markerów na twarzy.

2.1.4. Nagrania dźwiękowe

Nagrania dźwiękowe są zapisywane w postaci 16-kanałowych plików typu WAV z częstotliwością próbkowania wynoszącą 96 KHz i rozdzielczością bitową 16 bitów/próbkę. Na podstawie sygnałów dźwiękowych z 16 kanałów są generowane obrazy z kamery akustycznej. Obrazy powstają jako rezultat działania algorytmu adaptacyjnego kształtowania wiązki akustycznej. Obraz jest w istocie kwadratową macierzą rozkładu natężenia pola



Rysunek 2. Fragmenty przykładowych obrazów z kamery akustycznej. Obraz górny: wymowa głoski /e/, obraz dolny: wymowa głoski /n/

Rysunek 1. Przykładowe obrazy z kamer wideo. Kolejno od lewej są przedstawione obrazy z kamer: bocznej lewej, frontalnej, bocznej prawej

akustycznego, o wymiarach 50 × 50 punktów. Trójwymiarowa mapa rozkładu pola akustycznego nanoszona jest na poszczególne klatki obrazu z kamery wideo. Jeden piksel obrazu z kamery akustycznej reprezentuje obszar 5 x 5 mm rzeczywistego obrazu wideo. Po przeskalowaniu wynikowa rozdzielczość obrazu wynosi 1024 x 1024 piksele natomiast szybkość 200 klatek na sekundę, co jest zgodne częstotliwością ramek wideo oraz częstotliwością próbkowania sygnałów artykulograficznych. Przykładowe obrazy z kamery akustycznej przedstawiono na rysunku 2.

2.1.5. Pliki Text Grid

Pliki Text Grid zawierają informacje na temat segmentacji akustycznej nagrywanych słów. Są to pliki generowane przez program PRAAT [21]. Struktura przykładowego pliku jest pokazana w tabeli 1.

Pliki typu TextGrid zawierają informacje o chwilach początkowych i końcowych segmentów akustycznych. Segmentacja jest zapisana na dwóch poziomach. Poziom pierwszy (item[1]), zawiera informacje o chwili początkowej i końcowej całego słowa oraz o czasie trwania całego opisywanego nagrania dźwiękowego. Poziom drugi (item[2]) zawiera informacje na temat chwil początkowych i końcowych głosek wchodzących w skład słowa. Pliki TextGrid pełnią kluczową rolę w preyzyjnym wyszukiwaniu fragmentów nagrań reprezentujących poszczególne głoski w nagraniach audio, jak również w zgrubnym wyszukiwaniu analogicznych fragmentów w nagraniach wideo oraz artykulograficznych.

2.1.6. Lista wyrazowa

Lista wyrazowa obejmuje ok 500 słów języka polskiego, w których jest zawarty pełny inwentarz głosek polskich. Głoski z inwentarza występują w słowach w śródgłosie w akcentowanej pozycji. Słowa były dobierane w taki sposób aby głoski pojawiały się w jak najbardziej neutralnym kontekście. Spółgłoski występowały w słowach trójsylabowych w obustronnym kontekście samogłoski /a/ np. *ba**g**aże*, natomiast samogłoski występowały w słowach dwusylabowych w lewostronnym kontekście spółgłoski /p/ i prawostronnym kontekście spółgłoski /s/ np. *p**o**sag.*

2.1.7. Lista trifonów

Lista trifonów została stworzona w celu łatwego wyszukiwania słów z zadaną sekwencją 3 fonemów oraz w celu ułatwienia trenowania modeli sieci Bayesa. Trifon jest to symboliczny zapis głoski z uwzględnieniem lewego i prawego kontekstu. W bazie przyjęto zapis trifonów zgodny z konwencją pakietu oprogramowania HTK [22]. Lewy kontekst jest oznaczany znakiem '-', a prawy kontekst znakiem '+'. Przykładowo trifon zapisany jako m-a+g oznacza samogłoskę /a/, która jest poprzedzona spółgłoską /m/, i za którą stoi spółgłoska /g/. Specyficzny sposób zapisywania trifonów dotyczy głosek znajdujących się na początku

Tabela 1. Struktura przykładowego pliku TextGrid z segmentacją akustyczną

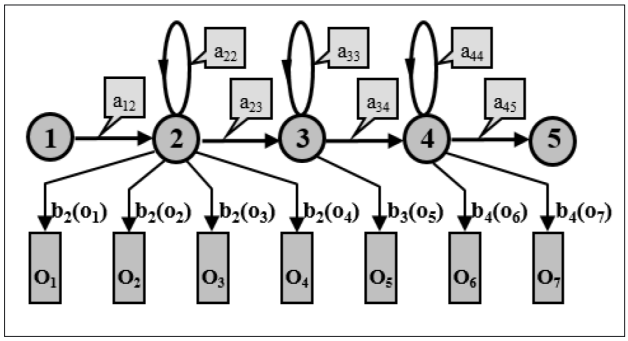
```
File type = "ooTextFile"
Object class = "TextGrid"

xmin = 0.000000
xmax = 5.589927
tiers? <exists>
size = 2
item []:
 item [1]:
 class = "IntervalTier"
 name = "1"
 xmin = 0.000000
 xmax = 5.589927
 intervals: size = 3
 intervals[1] :
 xmin = 0.000000
 xmax = 0.840010
 text = ""
 intervals[2] :
 xmin = 0.840010
 xmax = 1.305000
 text = "tlen"
 intervals[3] :
 xmin = 1.305000
 xmax = 5.589927
 text = ""
 item [2]:
 class = "IntervalTier"
 name = "2"
  xmin = 0.000000
  xmax = 5.589927
  intervals: size = 6
  intervals[1] :
  xmin = 0.000000
  xmax = 0.840010
  text = ""
  intervals[2] :
  xmin = 0.840010
  xmax = 0.984983
  text = "t"
  intervals[3] :
  xmin = 0.984983
  xmax = 1.055846
  text = "l"
  intervals[4] :
  xmin = 1.055846
  xmax = 1.142791
  text = "\ef"
  intervals[5] :
  xmin = 1.142791
  xmax = 1.305000
  text = "n"
  intervals[6] :
  xmin = 1.305000
  xmax = 5.589927
  text = ""
```

i na końcu wyrazu. Głoski te są zapisywane z uwzględnieniem tylko jednego kontekstu, np. trifon zapisany jako o+r oznacza samogłoskę /o/ na początku wyrazu (czyli poprzedzoną ciszą), za którą stoi spółgłoska /r/, natomiast trifon zapisany jako o-r oznacza spółgłoskę /r/ na końcu wyrazu (za którą stoi cisza) poprzedzoną samogłoską /o/.

2.1.8. Modele sieci Bayesa

Baza pozwala również na zapisywanie wytrenowanych modeli dynamicznych sieci Bayesa (ang. Dynamic Bayesian Network – DBN). Dynamiczne sieci Bayesa są uogólnieniem ukrytych modeli Markowa, a dość obszerny opis dynamicznych sieci Bayesa można znaleźć w literaturze [23]. W przypadku opisywanej bazy nagrań artykulograficznych dynamiczne sieci Bayesa mają służyć do doświadczeń w zakresie inwersji mowy. Baza POLEMAD będzie przystosowana do przechowywania dwóch rodzajów modeli dynamicznych sieci Bayesa: tradycyjnych ukrytych modeli Markowa (ang. Hidden Markov Model - HMM) oraz sprzężonych ukrytych modeli Markova (ang. Coupled Hidden Markov Model – CHMM). W obydwu modelach gęstość prawdopodobieństwa obserwacji będzie modelowana za pomocą mieszaniny wielowymiarowych rozkładów Gaussa (ang. Gaussian Mixture Model – GMM). Modelowanymi jednostkami będą fonemy. Wektorami obserwacji będą wektory otrzymane z sygnałów audio, wideo oraz artykulograficznych.

W przypadku HMM zasadniczo przewiduje się pięciostanowe modele Markowa dla fonemów pokazane na rys. 3.



Rysunek 3. Pięciostanowy ukryty model Markowa

W przypadku tego modelu zakładając mieszaninę *M* wielowymiarowych rozkładów Gaussa, gdzie prawdopodobieństwo obserwacji wyraża się wzorem:

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm}\, N(o_t, \mu_{jm}, \Sigma_{jm}) \qquad (1)$$

w bazie danych trzeba będzie przechowywać następujące parametry wytrenowanego modelu Markowa:

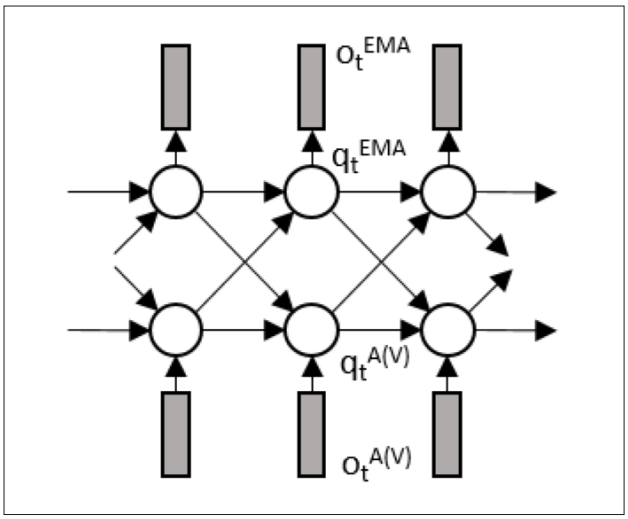
c_{2m}, c_{3m}, c_{4m} – wagi dla m-tego rozkładu Gaussa w mieszaninie rozkładów dla stanów 2, 3 i 4

μ_{2m}, μ_{3m}, μ_{4m} – średnie wektory obserwacji dla m-tego rozkładu Gaussa dla stanów 2, 3 i 4

$\sum_{2m}$, $\sum_{3m}$, $\sum_{4m}$ – macierze kowariancji dla m-tego rozkładu Gaussa dla stanów 2, 3 i 4.

A – macierz prawdopodobieństw przejść między stanami o wymiarze 5×5

W przypadku modeli CHMM przewiduje się model przedstawiony na rys. 4.



Rysunek 4. Struktura dwuwarstwowego sprzężonego ukrytego modelu Markowa

Do opisania modelu przedstawionego na rys. 4 trzeba będzie w bazie przechować zgodnie z [24] następujące parametry:

$P(q_t^s|\mathbf{q}_{t-1})$, s ∈{a(v),ema}: prawdopodobieństwa przejść między stanami

$P(\mathbf{o}_t^s|q_t^s)$, s ∈{a(v),ema}: prawdopodobieństwa obserwacji

Prawdopodobieństwa przejść między stanami będą zapisywane w postaci macierzy trójwymiarowej, natomiast prawdopodobieństwa obserwacji będą obliczane jako gęstość prawdopodobieństwa mieszaniny wielowymiarowych rozkładów Gaussa:

$$P(\boldsymbol{o}_t^s|q_t^s) = \sum_{k=1}^{K} w_{q_t^s k}\, N(\boldsymbol{o}_t^s, \mu_{q_t^s k}, \Sigma_{q_t^s k}) \qquad (2)$$

Gdzie:

$N(\boldsymbol{o}_t^s, \mu_{q_t^s k}, \Sigma_{q_t^s k})$ – k-ty wielowymiarowy rozkład Gaussa w mieszaninie dla obserwacji $\boldsymbol{o}_t^s$ dla stanu q_t^s,

$\mu_{q_t^s k}$ – średni wektor obserwacji dla k-tego wielowymiarowego rozkładu Gaussa w mieszaninie,

$\Sigma_{q_t^s k}$ – macierz kowariancji dla k-tego wielowymiarowego rozkładu Gaussa w mieszaninie

$w_{q_t^s k}$ – waga dla k-tego wielowymiarowego rozkładu Gaussa w mieszaninie

Zatem ze wzoru (1) wynika, że do wyliczenia prawdopodobieństw obserwacji dla stanu q_t^s w CHMM trzeba w bazie danych przechować średnie wektory obserwacji, macierze kowariancji oraz wagi dla wszystkich k wielowymiarowych rozkładów Gaussa w mieszaninie dla stanu q_t^s.

Ze względu na dużą liczbę wariantów zaprezentowanych powyżej modeli sieci Bayesa (różna liczba stanów, różna liczba

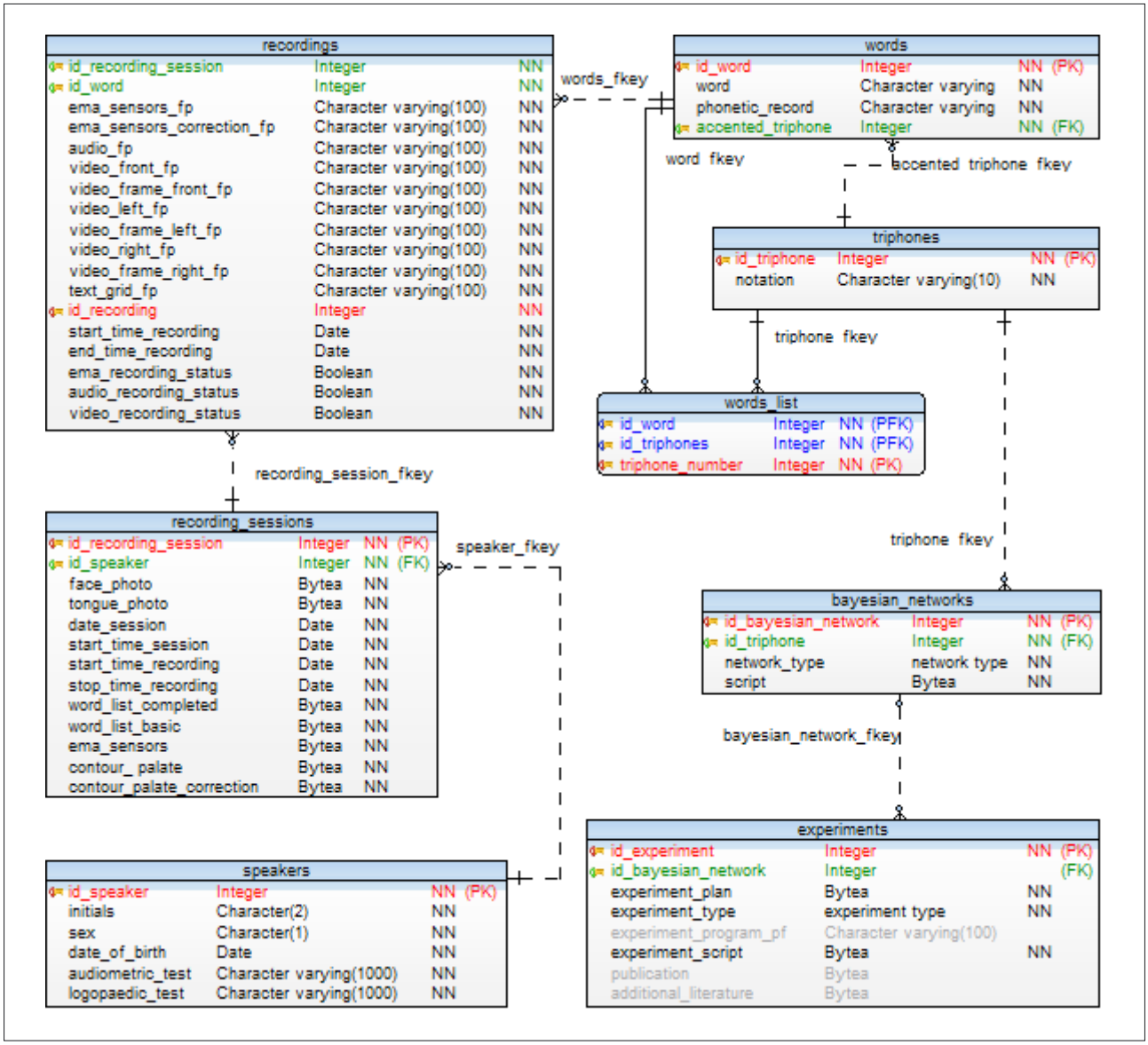
rozkładów w mieszaninie, różny rozmiar wektorów cech itp.) wszystkie parametry modeli dynamicznych sieci Bayesa (HMM oraz CHMM) będą zapisywane w pliku tekstowym w postaci skryptu w formacie podobnym do formatu HTK. Oprócz tego do pliku będzie dołączana lista fragmentów nagrań, na podstawie których były trenowane sieci Bayesa.

2.2. Struktura bazy

Jak już wspomniano baza danych będzie miała strukturę relacyjnej bazy danych. Przewiduje się osiem tabel, z których część pokrywa się bezpośrednio z rodzajami danych opisanymi w rozdziale 2.1. Tabele przedstawiono graficznie na diagramie ERD na rysunku 5. Klucz główny (PK) w tabeli wyróżniony jest czcionką koloru czerwonego, natomiast klucze obce (FK) kolorem zielonym. Czarny kolor czcionki oznacza, że atrybut nie może mieć wartości null (NN).

Poniżej zawarty jest opis poszczególnych tabel z rysunku 5.

1) Tabela *speakers* zawiera dane mówcy opisane w rozdziale 2.1.1. Klucz główny id_speaker.
2) Tabela *recording_sessions* zawiera dane sesji nagraniowej, w której uczestniczy dokładnie jeden mówca. Dane te obejmują: zdjęcie twarzy z markerami, zdjęcie języka z czujnikami EMA, datę sesji, chwilę początkową i końcową nagrań, listę wyrazową zrealizowaną (w odróżnieniu od listy wyrazowej pełnej), obrys podniebienia bez korekty ruchów głowy oraz ten sam obrys po korekcie. Klucz główny id_recording_session, klucz obcy id_speaker z tabeli *speakers.*
3) Tabela *recordings* Dane nagrań, czyli głównie nagrania pojedynczego słowa za pomocą artykulografu, kamery akustycznej oraz kamer wideo. Ze względu na duży rozmiar plików z nagraniami w bazie danych będą przechowywane ścieżki dostępu do plików, zamiast struktur danych zawierających te pliki. W tabeli znajdą się również ścieżki do



Rysunek 5. Diagram ERD bazy POLEMAD

plików TextGrid z segmentacją akustyczną słowa. Klucz główny id_recording, klucze obce id_recording_session z tabeli *recording_sessions* oraz id_word z tabeli *words*.

4) Tabela *words* zawierająca słowa z listy wyrazowej opisanej w rozdziale 2.1.6. Klucz główny id_word, klucz obcy accented_triphone z tabeli *triphones*.
5) Tabela *triphones* zawierająca kody trifonów z listy trifonów opisanej w rozdziale 2.1.7. Klucz głowny id_triphone.
6) Tabela *bayesian_networks* zawierająca opis dynamicznych modeli Bayesa (HMM, CHMM) w postaci skryptu (rozdział 2.1.8) Klucz głowny id_bayesian_network, klucz obcy id_triphone z tabeli *triphones*.
7) Tabela *experiments* zawierająca dane eksperymentów z wykorzystaniem danych z bazy, przede wszystkim dane wejściowe (materiał badawczy) i dane wyjściowe (wyniki), ale również opis eksperymentów oraz literaturę zawierającą podstawy teoretyczne. Eksperymenty są opisane w rozdziale 3. Klucz główny id_experiment, klucz obcy id_ bayesian_network z tabeli *bayesian_networks*.

Ponadto baza zawiera jedną tabelę pośredniczącą *word_list* ze względu na relację wiele do wielu między tabelą *words*, a tabelą *triphones*.

3. Planowane zastosowania bazy danych w eksperymentach naukowych

Opisywana baza danych ma z założenia służyć nie tylko do selektywnego pobierania informacji ale również ma ułatwiać wykonywanie eksperymentów naukowych. Główną zaletą korzystania z bazy danych podczas wykonywania eksperymentów jest stosunkowo łatwe, poprzez zapytania SQL, zautomatyzowane pobieranie dużej ilości danych przetwarzanych później za pomocą specjalistycznego oprogramowania oraz zapisywanie wyników eksperymentów w bazie. Przewidywane są dwa rodzaje eksperymentów, które będzie można wykonywać z użyciem bazy danych:

- Analizy statystyczne;
- Inwersja mowy.

Wykonywanie obydwu typów eksperymentów będzie odbywać się za pomocą skryptów. Skrypty eksperymentów mają za zadanie ułatwiać programom współpracującym z bazą danych generowanie zapytań SQL do pobierania danych z bazy oraz zapisywanie wyników eksperymentów w bazie. Obydwa rodzaje eksperymentów zostały bardziej szczegółowo przedstawione w kolejnych podrozdziałach.

3.1. Analizy Statystyczne

Analizy statystyczne obejmują badanie zmienności wewnątrzosobniczej [25] i międzyosobniczej [26] pomierzonych parametrów, testowanie istotności różnic między grupami jednostek fonetycznych [5] oraz badanie zależności korelacyjnych [27] regresyjnych [28].

Badanie zmienności wewnątrzosobniczej obejmuje wyznaczanie odchyleń standardowych (wariancji) lub innych miar rozproszenia parametrów artykulacyjnych, akustycznych lub video głosek dla danego mówcy. Może również obejmować wyznaczanie przedziałów ufności dla wspomnianych parametrów. Badanie zmienności międzyosobniczej może dotyczyć testowania istotności różnic między parametrami głosek między mężczyznami a kobietami. Może również obejmować testy istotności różnic ANOVA lub MANOVA w parametrach głosek między wieloma mówcami. Ze względu na ograniczoną liczbę przykładów głosek (trifonów) w bazie może zajść konieczność użycia testów nieparametrycznych do badania zmienności międzyosobniczej. Do testowania istotności różnic między grupami jednostek fonetycznych można użyć testów istotności różnic do prób zależnych. Przykładem testowania w próbach zależnych może być badanie różnic w położeniach czujników EMA dolnej i górnej wargi w fonemach /b/ i /p/ u tego samego mówcy. Badanie zależności korelacyjnych i regresyjnych odgrywa ważną rolę przy tworzeniu matematycznych modeli ruchu artykulatorów oraz w badaniach nad inwersją mowy. Szczególne znaczenie wydaje się mieć tutaj badanie regresji wielorakiej.

Liczba analiz statystycznych możliwych do wykonanie przy użyciu bazy danych wydaje się być bardzo duża. Przykładowo liczba współczynników korelacji dla jednej głoski między położeniami w osi X dla 8 czujników EMA, położeniami w osi X dla 39 markerów wideo oraz 50 parametrami akustycznymi sygnału dźwiękowego wynosi około 5000. Jeżeli liczbę tą przemnożyć przez 20 mówców, około 40 fonemów z inwentarza głosek języka polskiego oraz około 12 przykładów głosek to liczba współczynników korelacji wzrasta do około 50 milionów. Przykład ten pokazujący jedynie niewielką część możliwych do przeprowadzenia analiz uwidacznia konieczność automatyzacji obliczeń, co bez uporządkowania danych oraz relacji między nimi w postaci bazy danych jest zadaniem bardzo trudnym lub wręcz niemożliwym do zrealizowania.

3.2. Eksperymenty nad inwersją mowy

Planowane jest przeprowadzenie kompleksowych eksperymentów nad inwersją mowy. Inwersja mowy polega na odwzorowaniu jednego typu zapisanych sygnałów (obrazów) mowy na inny typ sygnału. Przykładem inwersji mowy może być odwzorowanie ruchu sensorów EMA umieszczonych na artykulatorach mowy w szczególności na języku na podstawie sygnałów akustycznych i/lub obrazów z kamery wideo. Przechowywane w opisywanej bazie dane umożliwiają następujące podstawowe rodzaje inwersji mowy:

- inwersja akustyczno-artykulograficzna – odtworzenie ruchu sensorów EMA na podstawie sygnałów akustycznych,
- inwersja wizyjno-artykulograficzna – odtworzenie ruchu sensorów EMA na podstawie sekwencji obrazów video,
- inwersja akustyczno-wizyjna – odtworzenie sekwencji obrazów wideo na podstawie sygnałów akustycznych.

Są oczywiście możliwe typy inwersji w odwrotnym kierunku np. inwersja artykulograficzno-akustyczna, w której sygnał akustyczny jest syntezowany na podstawie sekwencji położeń czujników EMA. Są również możliwe typy kombinowane np. inwersja akustyczno-wizyjno-artykulograficzna, w której ruchy artykulatorów są odtwarzane na podstawie sygnałów audio oraz obrazów wideo, w tym przypadku połączenie dwóch typów nagrań może potencjalnie przynieść korzyści w postaci dokładniejszego odwzorowania ruchu czujników EMA.

Do przeprowadzania inwersji mowy stosuje się różnorodne metody, których przegląd można znaleźć w pracy [29]. W przypadku bazy POLEMAD przewiduje się przeprowadzenie doświadczeń nad inwersją mowy z wykorzystaniem metody DTW [30] oraz dynamicznych sieci Bayesa [24]. Tak jak wspomniano w rozdziale 2.1 jest planowane użycie 2 rodzajów sieci Bayesa: Ukrytych modeli Markowa (HMM) oraz sprzężonych ukrytych modeli Markowa (CHMM). W przypadku metody DTW do doświadczenia są potrzebne surowe dane z nagrań np. nagranie dźwiękowe oraz zapis położeń czujników EMA. Natomiast do doświadczeń z wykorzystaniem ukrytych modeli Markowa oraz dynamicznych sieci Bayesa konieczne jest skorzystanie z wytrenowanych modeli zapisanych w bazie danych.

4. Podsumowanie

W artykule opisano strukturę bazy danych do przechowywania danych z badań artykulograficznych. Najważniejszymi rodzajami danych przechowywanych w bazie danych są nagrania położeń i kątów nachyleń czujników artykulografu, nagrania z trzech kamer wideo oraz nagrania z 16-kanałowego rejestratora audio. Obecnie baza ma zaprojektowaną strukturę tabel. Trwają prace nad migracją nagrań 20 mówców do bazy danych. Informacje zapisane w bazie pozwolą na przeprowadzanie kompleksowych analiz statystycznych oraz doświadczeń nad inwersją mowy. Planowane jest udostępnienie danych z bazy w marcu 2018 roku. Baza będzie dostępna pod adresem: http://polemad.pwsz.tarnow.pl.

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.

Bibliografia

1. J. S. Perkell, M. H. Cohen, M. A. Svirsky, M.L. Matthies, I. Garabieta, and M. T. Jackson, *The Journal of the Acoustic Society of America*, 1992, **92(6)**, 3078–3096.
2. H. Kjellström, O. Engwall, Audiovisual to articulatory inversion, *Speech Communication*, 2009, **51(3)**, 195–209.
3. A. Katsamanis, G. Papandreou, and P. Maragos, *Audiovisual-to-Articulatory Inversion Using Hidden Markov Models*, Proceedings of the IEEE Workshop on Multimedia Signal Processing (MMSP-2007), 2007, 457–460.
4. K. Richmond, *Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus*, Proceedings of 12th Annual Conference of the International Speech Communication Association INTERSPEECH 2011, 1505–1508.
5. A. Lorenc, *Wymowa normatywna polskich samogłosek nosowych i spółgłoski bocznej*, Dom wydawniczy ELIPSA, Warszawa 2016, ISBN 978-83-8017-090-2.
6. D. Król, A. Lorenc, „Tarnowskie Colloquia Naukowe”, 2017, **4(3/2017)**, 9–16.
7. MOCHA-TIMIT database (2001), available online, http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html, (accessed December 2017).
8. S. Narayanan, A. Toutios, V. Ramanarayanan, et al., *The Journal of the Acoustical Society of America*, 2014, **136(3)**, 1307–1311.
9. F. Rudzicz, A. K. Namasivayam, T. Wolff, *Lang Resources & Evaluation,* 2012*,* **46**, 523–541.
10. A. Ji, J. J. Berry, M. T. Johnson, *The Electromagnetic Articulography Mandarin Accented English (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data*, 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,* 2014, 7719–7723.
11. J. Beskow, O. Engwall, and B. Granström, *Resynthesis of Facial and Intraoral Articulation fromSimultaneous Measurements*, Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS’03), 2003.
12. E. Meister, L. Meister, *Multimodal Corpus of Speech Production: Work in Progress*, in book: *Human Language Technologies. The Baltic Perspective: Proceedings of the Fifth International Conference Baltic HLT 201*2, Edition: Frontiers in Artificial Intelligence and Applications, IOS Press, 2012, ch. Multimodal Corpus of Speech Production: Work in Progress, pp.146–153.
13. M. Rochoń, B. Pompino-Marschall, *The Articulation of Secondarily Palatalized Coronals in Polish. In Proceedings of XIVth International Congress of Phonetic Sciences*, San Francisco, 1999, 1897–1900.
14. B. Pompino-Marschall, M. Żygis, *Surface Palatalization of Polish Bilabial Stops: Articulation and Acoustics. Proce-*

edings of the 15th International Congress of Phonetic Sciences, 2003, 1751–1754.
15. A. Trochymiuk, R. Święciński, *Logopedia*, 2009, **38**, 173–201.
16. A. Lorenc, R. Święciński, *Application of Phonetics in Speech Therapy: a Case of Abnormal Convex Tongue Setting in Polish. in Recent Developmnets in Applied Phonetics. Studies in Linguistics and Methodology 6*, Wydawnictwo KUL, Lublin, 2014, 287–324.
17. R. Święciński, *An EMA Study of Articulatory Settings in Polish Speakers of English. in Teaching and Researching English Accents in Native and Non-native Speakers*, Springer, Heidelberg, 2013, 73–82.
18. Ł. Mik, R. Wielgat, A. Lorenc, D. Król, R. Święciński, R. Jędryka, *Multimodal Speech Data Acquisition with the Use of EMA Fast-speed Video Cameras and a Dedicated Microphone Array, Proceedings of 2016 MIXDES – 23rd International Conference Mixed Design of Integrated Circuits and Systems*, 2016, 415– 418.
19. P. Hoole and A. Zierdt, *Five-dimensional articulography, Speech Motor Control: New developments in basic and applied research*, eds. B. Maassen and P.H.H.M. Van Lieshout, 2009, 331–349.
20. M. Stella, P. Bernardini, F. Sigona, A. Stella, M. Grimaldi, B. Gili Fivela, *J. Acoust. Soc. Am.*, 2012, **132(6)**, 3941– 949.
21. P. Boersma, D. Weenink, „Praat: doing phonetics by computer" [computer program, version 5.3.57]. webpage: http://www.praat.org/, 2014.
22. Hidden Markov Model Toolkit (HTK), available online, http://htk.eng.cam.ac.uk/, (accessed December 2017).
23. K. Murphy, *Dynamic Bayesian networks: Representation, inference and learning*, Ph.D. thesis, UC Berkeley, Computer Science Division (2002).
24. Xie, L., Liu, Z.-Q., *Pattern Recognition*, 2007, **40(8)**, 2325–2340.
25. A. Lorenc, R. Wielgat, „Tarnowskie Colloquia Naukowe" – *Nauki humanistyczne*, 2017, **(2)1/2017**, 129–157.
26. R. Wielgat, A. Lorenc, *Science, Technology and Innovation*, 2017, zgłoszone do publikacji.
27. R. Wielgat, Ł. Mik, A. Lorenc, A. Truchan, M. Szostek, *Choice of optimal measurement conditions for calculating the correlation between EMA sensor and video marker position coordinates in electromagnetic articulography, Proceedings of 2017 International Conference on Systems*, Signals and Image Processing (IWSSIP), Poznań, 2017.
28. R. Wielgat, Ł. Mik, A. Lorenc, *Correlational and regressive analysis of the relationship between tongue and lips motion – An EMA and video study of selected polish speech sounds, 2017 MIXDES – 24th International Conference "Mixed Design of Integrated Circuits and Systems"*, 509-514, 2017.
29. A. Ji, *Speaker Independent Acoustic-to-Articulatory Inversion*, Dissertation, Marquette University, 2014.
30. R. Wielgat, A. Lorenc, *Speech inversion by dynamic time warping method, 2016 International Conference on Signals and Electronic Systems (ICSES)*, 2016, 81–84.

Article history:
Received 1 December 2017
Received in revised form
22 December 2017
Accepted 23 December 2017
Available online 27 December 2017

Abstract
The article describes the structure and functionality of the articulographic database for storing data from articulographic research using an electromagnetic articulograph, an acoustic camera and 3 video cameras. The database enables selective extraction of various types of data for scientific research and interoperates with programs that carry out experiments. Structure and construction of the database is described. Potential future application in statistical analysis and experiments on speech inversion using dynamic Bayesian networks (DBN) was also presented.

Key words: electromagnetic articulography, database, Bayesian networks, speech inversion, acoustic camera, articulatory phonetics, acoustic phonetics