In: Eugeniusz Cyran and Jolanta Szpyra-Kozłowska (Eds) *Crossing Phonetics-Phonology Lines*. Newcastle Upon Tyne: Cambridge Scholars Publishing. 2014. Pp. 299-312.

ARTICULATORY GROUNDING OF PHONEMIC DISTINCTIONS IN ENGLISH BY MEANS OF ELECTROPALATOGRAPHY

GRZEGORZ KRYNICKI

1. Introduction

The aim of the experiment described in this paper was to devise and test a procedure that would allow identification of a phoneme on the basis of only tongue-to-palate and labial (EPGL) contacts that accompanied its realization in continuous read speech. The hypothesis underlying this study was that the articulatory correlates of the phonemic distinctive features, despite the unstable character of these correlates in comparison to acoustic realisations, can be induced statistically from dimensionality-reduced EPGL data. Most phoneme recognition studies so far have used posterior probabilities of symbolic articulatory features rather than physical data as observations in their classifiers and none has used EPG as the only source of data.

The system used to obtain EPGL data was the CompleteSpeech Palatometer (CompleteSpeech 2013). The system was used on a single subject, a 24-year-old female speaker of General American, reading sentences that used specific phonemes at the same frequency they appear in English (see Section 3) plus the alphabet and numerals. The recordings were segmented and annotated with phoneme labels using Penn Forced Aligner (Jiahong and Liberman 2008). The EPGL information was transformed into a set of linguistically meaningful and computationally manageable parameters – dimensionality reduction indices (DRI) using modified techniques developed by Hardcastle et al. (1991). 11 DRI's were calculated for each EPGL matrix: the number of electrode activations in the alveolar (ALV), palatal (PAL) and velar regions (VEL), total number of contacts (TOT), centre of gravity (COG), posterior centre of gravity (POS), anterior centre of gravity (ANT), laterality (LAT), asymmetry (ASY), frication (FRI) and labiality (LAB).

Two classification methods were used to predict the phonemic category of each token based on its EPGL parameters: forward-selection linear discriminant analysis (LDA) and a probabilistic neural network (PNN). Better classification results were obtained by PNN model (32.1%). LDA made it possible to establish which DRI's had a significant influence on the classification result in the decreasing order of the influence: VEL TOT PAL ALV FRI POS BIL ASY COG LAT ANT. Average pair-wise phoneme LDA classification rate was 88.2%.

2. Previous research

In previous studies of articulatory-based phoneme recognition or automatic speech recognition, one of two approaches has been adopted. Most commonly, pseudo-articulatory information in the form of symbolic features (voiced, fricative, etc.) has been used as the basis for classification (Kirchhoff 1999, Kirchhoff et al. 2002, Deng and Sun 1994, Metze and Waibel 2002, c.f. Bates et al. 2007: 84). Less frequently, physically recorded articulatory data or parameters derived from that data were used (X-ray in Blackburn and Young 1996, electro-magnetic articulography in Fagan et al. 2008, EMA + electroglottograph + EPG in Uraga and Hain 2006). The author is not aware of any study in which phoneme recognition has been conducted based on EPG data only.

Uraga and Hain 2006 is the only study that used EPG to ASR/phoneme recognition with other articulatory information, specifically EMA and EGG. The phone error rate they report was 35.7%. Information derived from EPG yielded only 1.5% error reduction.

Another attempt to use a so-called electropalatogram for ASR was made by Jorgensen 2003. A classifier was trained to recognize 6 sub-vocally pronounced discrete words based on the EMG/EPG signal that accompanied their pronunciation. However, the understanding of electropalatogram adopted by the authors differs from the usual understanding of the term, namely it was meant as "EMG measured [...] under the chin to pick up surface tongue signals". No tongue-to-palate contacts were directly measured.

A related process of estimating electropalatographic patterns from the speech signal as a case of speech inversion was described in Toutios 2008. EPG patterns from a single speaker were reconstructed from the estimated projections on 9 components obtained through Principal Component Analysis (PCA). PCA components were extracted statistically with no reference to phonological oppositions (anterior-posterior, etc.). In

that study, binary EPG patterns were reconstructed from the estimated projections on principal components with the error rates from 22.34% to just 2.67% depending on the number of components in the model. Phone classification into phoneme categories based on these components was not attempted.

EPG information alone has not been used as the only basis of ASR or phone recognition for several reasons. First, it provides an incomplete articulatory description: it does not provide information about nasalisation, voicing, lip position and tongue position when no contact to the palate is made. On the other hand, there are methods that can provide such information if necessary. Second, interspeaker variability in terms of articulation (due to different palate sizes and shapes or other speaker-specific physical characteristics) appears to be greater than inter-speaker variability in acoustic realisation of phonemes (Pierrehumbert 2000: 6). Third, the "same" sounds can be created by a single speaker using a range of different articulatory gestures (Neiberg et al. 2008), which is particularly true of back vocoids (Ladefoged 2006: 189, Lodge 2009: 42). Finally, easy access to acoustic information and the relatively difficult access to articulatory information made articulatory speech and phoneme recognition of relatively low practical value.

It is believed here however that recent advancements in the area of articulatory tracking will result in non-invasive and affordable artificial palates for accurate imaging of not just tongue-to-palate contacts but also movements of the tongue, lips and even soft palate (Wrench et al. 1996, Birkholz and Neuschaefer-Rube 2011, 2012). Such technology combined with research on articulatory variability and phone-, word- and sentence-recognition algorithms could significantly improve robustness of automatic speech recognition when speech is masked by background noise (Kirchhoff 1999, Mitra et al. 2011), enable oral communication in silence-restricted environment (e.g. military or security operations, Hueber et al. 2010) and provide laryngectomized patients with a more efficient and natural mode of oral communication (Wang et al. 2012). It could also help in L1 speech therapy or L2 pronunciation training by providing feedback on patient's or student's phoneme recognition rate relative to the recognition rates of correct or native models.

3. The data

The SmartPalate Palatometer comes with an artificial palate consisting of a custom-made flexible 0.5mm layer of biocompatible acrylic. The palate is embedded with 124 electrodes recording tongue-to-palate contacts and

2 electrodes recording labial contacts. The arrangement of the electrodes does not vary with different palate shapes and sizes. The system offers the largest number of electrodes of the systems available on the market (as of 2013). It is also the only system that displays labial contacts. The contacts are recorded at 100Hz frequency and may be viewed in real time on the computer screen, aligned with audio, saved and exported for further analysis in time-aligned WAV and TXT formats.

SmartPalate has not been used in as many studies as other EPG solutions, possibly because of, on the one hand, its relatively recent release and, on the other, because of the grid electrode layout (c.f. Wrench 2012). In this layout, each electrode is positioned by a fixed distance from the neighbouring electrodes, which would usually require a non-trivial step of software between-speaker normalization before pronunciations of different speakers are comparable. In manually normalized layout, each electrode is aligned with anatomical features and generally does not require software normalization.

In the few studies where SmartPalate was used (Pantelemidou et al. 2003, Schmidt 2007), it was found to provide accurate and consistent visual feedback for speech therapy. In this study, since the data were obtained from a single speaker, EPG matrices generated by the grid electrode layout did not need to be normalized between speakers.

The speaker read a list of 197 items: 10 sentences from the List 11 of Harvard Sentences (Harvard Sentences 2013), 130 phonetically balanced sentences from the TIMIT prompt list (TIMIT 2013), English alphabet and numerals from 0 to 30. The prompts were adjusted to provide for fillers, re-starts and minor misreadings on the part of the speaker. The whole list contained 1254 words.

The acoustic data were saved in PalateView as a WAV file annotated internally for EPG information. EPG annotation was exported to a text file in PalateView and removed from WAV by file conversion in Audacity. Then the prompts were phonemically annotated and force-aligned with the WAV file by means of Penn Forced Aligner. The output of force-alignment was formatted as a Praat TextGrid. The output was not corrected manually. The difference between phone boundaries in data aligned by means of Penn Forced Aligner and the gold standard was estimated by Gorman et al. (2011) at the mean of 0.2061 sec. and the median of 0.0124 sec., which has been considered not fatal for the requirements of this study. As such, the database contained 16 min. of audio data, 1254 words and 5303 segments, each labelled with one of the

phonemes from the set adopted by Penn Forced Aligner after CMU Pronouncing dictionary (The CMU Pronouncing Dictionary 2013).

3.1. Time alignment of phonemic and EPGL data

Time alignment between phonetic and EPGL data consisted in matching the time point in the middle of each phone and obtaining EPGL data available for the corresponding time point. In order to compensate for the potential asynchrony between articulation and acoustic signal, EPGL data for different time values around the mid-time were tried to maximise the categorisation into phonemic classes (to be discussed in Section 4). These values were time points around the middle of each phone segment (identified by Penn Forced Aligner and tagged phonemically) from -100ms to 200ms every 10ms. To obtain a single classification result, the same time offset was applied to all phones irrespective of their duration. Classification was conducted using two different classifiers: PNN and LDA. The results presented in Fig. 1 show that the overall most successful classifier (PNN) achieved its best result (32.1%) when trained on the EPGL image of a phone from the time point that was nearest to the middle of the acoustic image of that phone.

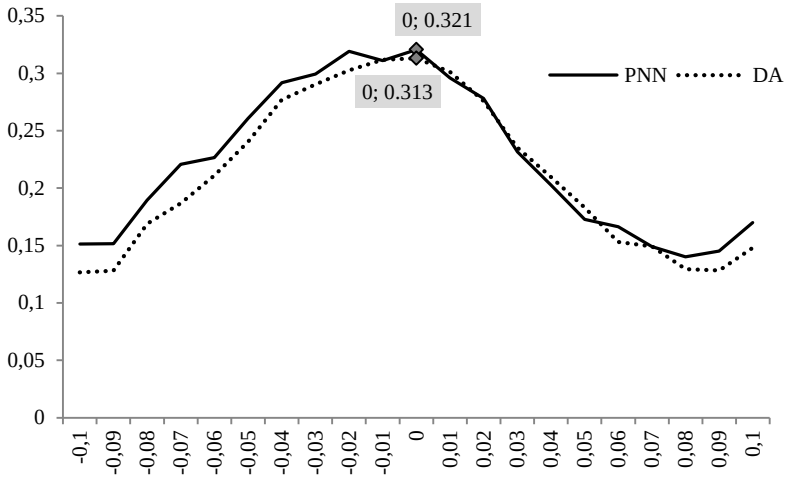

Fig. 1 Results for Linear Discriminant Analysis and Probabilistic Neural Network classifiers depending on time point for which EPGL data were obtained relative to the middle of the classified acoustic phone.

Both classifiers performed best for EPGL data obtained for time points synchronous with corresponding data points of the acoustic signal. Time

alignment between phonetic and EPGL data was therefore performed without any time offset.

3.2. Parametrisation of the EPGL data

Parameterisation of EPG data is performed to make the data more amenable to analysis, since direct manipulation of the raw EPG sequences is difficult due to its high dimensionality. Two approaches to reduce EPG matrixes into computationally manageable parameters are generally applied: linguistically meaningful dimensionality reduction indexes (DRI) based on binary data methods as collected and reviewed in Hardcastle et al (1991) and probabilistic data-driven indexes in the form of PCA components (Nguyen et al. 1996, Carreira-Perpiñán and Renals 1998). The latter approach has no linguistic assumptions and its results are not easy to interpret. Because one of the aims of this experiment was to find articulatory correlates of phonemic distinctive features, the linguistically-motivated approach to data reduction was chosen.

Each of DRI's was calculated for each EPGL data point that was time-aligned with the corresponding phone. Calculations were conducted based on Hardcastle 1991: 251-266, Harrington 2011: 241-243 and Carreira-Perpiñán and Renals 1998: 262.

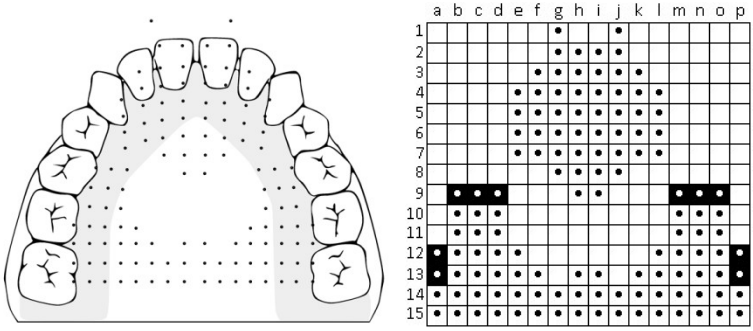

Fig. 2 Map of EPGL electrodes in the SmartPalate device (left) and its schema (right): black dots represent electrodes activated at least once in the experiment, white dots represent electrodes not activated even once.

BIL – (bilabial) sum of contacts in 1st row divided by all electrodes in that row

ALV – (alveolar) sum of contacts in rows 2-4 divided by all electrodes in these rows

PAL – (palatal) sum of contacts in rows 5-9 divided by all electrodes in these rows

VEL – (velar) sum of contacts in rows 10-15 divided by all electrodes in these rows

TOT – (total) sum of contacts in rows 1-15 divided by all electrodes in these rows

COG – (general centre of gravity) weighted average of the sum of contacts in rows 1-15, where the weights on rows are 14, 13, ..., 1

ANT – (anterior centre of gravity) weighted average of the sum of contacts in rows 2-8 columns G-J, where the weights on rows are 14, 13, ..., 8

POS – (posterior centre of gravity) weighted average of the sum of contacts in rows 9-15 columns D-M, where the weights on rows are 7, 6, ..., 1

LAT – (laterality) weighted average of the sum of contacts in rows 1-15 columns D-M, where the weights on columns are 1, 2, ..., 8 in columns A-H and 8, 7, ...,1 in columns I-P

ASY – (asymmetry) difference between the sum of contacts in columns A-H and I-P

FRI – (fricativity) the sum of contacts in rows 2-5 divided by all electrodes in these rows minus the sum of contacts in rows 6-8 divided by all electrodes in these rows

4. Classification procedure and results

Classifiers used to predict the phonemic class based on its EPG parameters were PNN (implemented in Statgraphics) and LDA (for classification of all phonemes Statistica was used, for classification of phoneme pairs R statistical package was used). 3210 unique phoneme-DRI pairs were used to develop classification models that discriminated among 38 phonemes; the /ʒ/ phoneme was excluded from all classifications as it was illustrated by only 2 cases. In the case of both LDA and PNN, jack-knifing (leave-one-out) was used as a cross-validation technique. Prior probabilities used were proportional to observed.

Time point	PNN	LDA
-0.03	29.9	29.0
-0.02	31.9	30.2
-0.01	31.1	31.2
0	32.1	31.3

Table 1 Results for 3 classification methods for different time-points

The results of the classification for 4 time points and 3 classifiers are presented in Table 1. The best classifier trained on all phonemes was PNN and it performed with 32.1% success rate. Relative similarity of the results of LDA (31.3%) suggests that the overall low correct classification rate is not a result of applying linear classification methods to data that are not linearly separable. Rather, it may indicate that the training set was too small considering the complexity of the task and that the phoneme classification problem cannot be easily solved without considering the probabilities of certain phoneme or word sequences (e.g. in the form of Hidden Markov models and N-gram models, c.f. Uraga and Hain 2006).

LDA classification method allows the possibility of estimating prediction power of EPGL parameters. Table 2 presents the relative weight associated with the first linear discriminant function that best separates the majority of cases.

DRI	StdCoeff	abs(StdCoeff)
VEL	4.25788	4.25788
TOT	-3.80678	3.80678
PAL	2.2047	2.2047
ALV	1.38722	1.38722
FRI	0.531696	0.531696
POS	0.172829	0.172829
BIL	0.132589	0.132589
ASY	-0.07888	0.078883
COG	-0.02624	0.02624
LAT	0.011153	0.011153
ANT	0.000772	0.000772

Table 2. Standardized coefficients for the first discriminant function. The higher the absolute standardized coefficient, the more a given DRI contributes to discriminant function.

The above result shows that velarity (as defined in Section 3.2) has the strongest associated weight with the first linear discriminant function and therefore has the greatest prediction power for discriminating between different phones based on their EPGL images only. Considering a limited character of the acoustic database that constituted the basis for this study, this result may indicate that the most reliable articulatory correlate of phonemic distinctions in American English is velarity. The second most reliable predictor is the total number of tongue-to-palate contacts. Anterior centre of gravity was found to be the least important contributor to the first discriminant function.

In the final step of the analysis, pairwise comparisons were conducted for all 692 combinations of 38 phonemes (all 703 combinations minus 11 whose EPGL data were of insufficient variability for the classification). The average correct classification rate for pairwise classification of 692 combinations was 88.2%.

Pairwise classification can be illustrated by the classification of 256 cases of /s/ vs. /ʃ/ phonemes. All 11 DRI's were entered into the LDA model as predictor variables. The LDA model correctly classified 87.89% of the cases. The first discriminating function was:

*SDF = - 16.0075*vel + 14.0958*tot - 5.8473*pal -2.8013*alv + 1.19524*lat - 1.0486*fri - 0.8856*asy - 0.2584*ant + 0.2559*pos - 0.0818*cog - 0.0645*bil*

From the relative magnitude of the standardized discriminant function coefficients in the above equation, it can be determined how DRI's are used to discriminate amongst the two phonemes. It can be seen that velarity (*vel*) as defined in 3.2 above is the most reliable predictor of the /s/-/ʃ/ variable, however other predictors also play a significant role in discriminating between these phonemes, in particular the total number (*tot*) of the contacts and contacts in palatal (*pal*) and alveolar (*alv*) region. Other features incl. laterality and bilabiality play a negligible role in differentiating between phonemes /s/ and /ʃ/.

Results of classification of each pair of phonemes together with LDA coefficients used to estimate the relative importance of predictors in discriminating between phonemes in each of these pairs can be accessed in Krynicki 2013.

5. Applications and implications

The information about articulatory correlates of pairwise phoneme distinctive features may be relevant in speech therapy and L2 pronunciation learning. A plausible scenario would include a student equipped with an EPGL device, trying to master pronunciation differences between two phonemes, e.g. /s/ and /ʃ/, as illustrated by their respective model articulations. In EPGL matrices of model and student's articulations, special software could highlight the areas that play a key role in differentiating between these phonemes in the speech of a native speaker, in our example that would be contacts in rows 10-15 (velar).

Further, learner and model articulations could be parameterised to continuous multilinear representations (Fig. 3) and compared, providing feedback to the learner on the parameters that differed the most. All critical articulatory gestures (see King and Taylor 2000) could be presented to the learner in the multilinear representation conveniently illustrating and explaining the phenomena of assimilation and coarticulation.

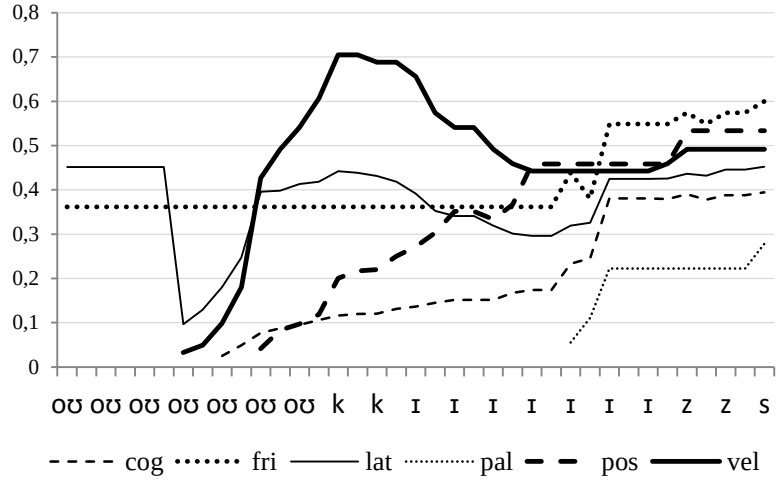

Fig. 3 Multilinear representation of the utterance [ʊʊk ɪz s] from the reading of *Oak is strong and also gives shade.* Some DRI's were removed for clarity.

Multilinear representation of parameters extracted from continuous EPGL data stream as presented in this study is analogical to tiers of phonetic features used in Browman and Goldstein (1989). Both are based

on the assumption that articulatory gestures constitute important phonological units and that they more adequately represent acoustic events than the phoneme. In our approach, the phoneme corresponds to a combination of articulatory features defined in terms of DRI's, represented on different tiers and overlapping in time. These features need not be binary and they may vary in their influence on the articulatory distinctiveness of the phones they constitute. It is argued here that such representation can be effectively used to teach L1 and L2 pronunciation, provide insights in the study of assimilation and coarticulation processes and provide predictions for missing information in the acoustic signal in robust automatic speech recognition and silent speech interfaces (c.f. Deng and Sun 1994, Carson-Berndsen 2000, Kane and Carson-Berndsen 2011)

6. Summary

The experiment described in the paper consisted in the classification of phones into phoneme categories on the basis of articulatory features obtained from electropalatographic and labial data. Best classification results were obtained with EPGL and acoustic data aligned with no time offset and they were produced by PNN (32.1%). Velarity was found to be the most reliable articulatory correlate of phonemic distinctions. In the experiment that consisted in classification of pairs of phonemes, only LDA was used and the average classification result was 88.2%.

It is understood, however, that the classification of segmented and static EPGL matrixes of phonemes is only the first step in the processing of continuous stream of EPGL matrixes which is not discrete and which involves complex co-articulatory processes. Future work will focus on segmentation of the continuous EPGL signal, fine-tuning of EPGL parameters, obtaining more data from greater number of speakers and combining the classification results with models of phoneme- and word-sequence probabilities.

References

Bates, R., M. Ostendorf - R. Wright (2007) Symbolic phonetic features for modeling of pronunciation variation. *Speech Communication* 49(2): 83-97.

Bates, R., M. Ostendorf - R. Wright (2007) Symbolic phonetic features for modeling of pronunciation variation. *Speech Communication* 49(2): 83-97.

Blackburn, C. S. - Steve J. Young (1996) Pseudo-Articulatory Speech Synthesis For Recognition Using Automatic Feature Extraction From X-Ray Data. ICSLP 1996 v.2, volume 2, pages 969-972.

Browman, C.P. - Goldstein, L. (1989) Articulatory gestures as phonological units. In: Phonology 6, Cambridge University Press, Cambridge: 201–251.

Carson-Berndsen, Julie - Michael Walsh. 2006. Phonetic Time Maps. In: *Text speech and language technology*, Nancy Ide and Jean Véronis (Eds). Springer: Dordrecht, 45-66.

Carreira-Perpiñán, M. - S. Renals (1998) Dimensionality reduction of electropalatographic data using latent variable models. *Speech Communication* 26(4): 259-282.

Deng, L. and Sun, D.X., 1994. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. J. Acoust. Soc. Am. Volume 95, Issue 5, pp. 2702-2719.

The CMU Pronouncing Dictionary, URL: http://www.speech.cs.cmu.edu/cgi-bin/cmudict, Last access: 30-09-2013.

CompleteSpeech, URL: http://completespeech.com, Last access: 30-09-2013.

Fagan, M. J., S. R. Ell, J. M. Gilbert, E. Sarrazin - P. M. Chapman (2008) Development of a (silent) speech recognition system for patients following laryngectomy, *Medical Engineering & Physics* 30(4): 419–425.

Goldsmith, J. (1990) Autosegmental and Metrical Phonology. Basil Blackwell, Cambridge, MA.

Gorman, K., J. Howell - M. Wagner (2011) Prosodylab-Aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics* 39(3): 192–193.

Hardcastle, W.J., F.E. Gibbon - K. Nicolaidis (1991) EPG data reduction methods and their implications for studies of lingual coarticulation. *Journal of Phonetics* 19: 251-266.

Harrington, J. (2010) *Phonetic Analysis of Speech Corpora.* Malden – Oxford: John Wiley & Sons.

Harvard Sentences (2013) URL: http://www.cs.columbia.edu/~hgs/audio/harvard.html, Last access: 30-09-2013.

Hueber, Thomas - Elie-Laurent Benaroya - Gerard Chollet - Bruce Denby - Gerard Dreyfus - Maureen Stone (2010) Development of a

silent speech interface driven by ultrasound and optical images of the tongue and lips. Speech Communication 52(4), 288-300.

Jiahong, Y. - M. Liberman (2008) Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics* 2008: 5687-5690.

Kane, Mark - Julie Carson-Berndsen. 2011. Multiple source phoneme recognition aided by articulatory features. Trends in Applied Intelligent Systems. Lecture Notes in Computer Science 2011, Volume 6704/2011 - IEA/AIE, 426-435

King, Simon - Paul Taylor. (2000) Detection of Phonological Features in Continuous Speech using Neural Networks. Computer Speech and Language. 14(4): 333-353.

Kirchhoff, Katrin (1999) Robust Speech Recognition Using Articulatory Information, PhD Thesis, University of Bielefeld, Germany.

Kirchhoff, Katrin - Gernot A. Fink - Gerhard Sagerer (2002) Combining acoustic and articulatory feature information for robust speech recognition. Speech Communication. Volume 37, Issues 3–4, July 2002, Pages 303–319.

Krynicki, Grzegorz (2013) Table of pair-wise phoneme classifications by means of LDA. URL: http://wa.amu.edu.pl/~krynicki/pub/apap, Last access: 18-03-2014

Ladefoged, P. (2006) A course in phonetics. (5th edition) Boston, MA: Thomson Wadsworth.

Lodge, Ken (2009) Critical Introduction to Phonetics. London: Continuum.

Metze, F. - A. Waibel (2002) A flexible stream architecture for ASR using articulatory features. *Proceedings of ICSLP*: 2133–2136.

Mitra, Vikramjit - Hosung Nam - Carol Y. Espy-Wilson (2011) Robust speech recognition using articulatory gestures in a Dynamic Bayesian Network framework. Proc. of Automatic Speech Recognition & Understanding Workshop, ASRU, 131-136.

Neiberg, Daniel - G. Ananthakrishnan - Olov Engwall. 2008. The Acoustic to Articulation Mapping: Non-linear or Non-unique? INTERSPEECH 2008, 1485-1488.

Nguyen, N., A. Marchal - A. Content (1996) Modeling tongue-palate contact patterns in the production of speech. Journal of Phonetics, 24, 1, 77-98.

Pantelemidou, V., R. Herman - J. Thomas (2003) Efficacy of speech intervention using electropalatography with a cochlear implant user. *Clinical Linguistics & Phonetics* 17(4/5), 383-392.

Pierrehumbert, J. (2000) The phonetic grounding of phonology. *Les Cahiers de l'ICP, Bulletin de la Communication Parlée* 5: 7-23.

Schmidt, A. M. (2007) Evaluating a new clinical palatometry system. *Advances in Speech–Language Pathology* 9(1): 73-81.

TIMIT (2013) URL: http://web.mit.edu/course/6/6.863/share/nltk_lite/timit/sentences, Last access: 30-09-2013.

Toutios, A. - K. Margaritis (2008) Estimating electropalatographic patterns from the speech signal. *Computer Speech & Language* 22: 346-359.

Uraga, E. - T. Hain (2006) Automatic Speech Recognition Experiments with Articulatory Data. *Proceedings of ICSLP*: 353-356.

VowelViz. URL: www.completespeech.com/speech/vowelviz1 Last access: 18-03-2014

Wang, Jun - Ashok Samal - Jordan R. Green - Frank Rudzicz. 2012. Sentence recognition from articulatory movements for silent speech interfaces. ICASSP Proceedings. 4985-4988.

Wrench, A.A. - A.D. McIntosh - W.J. Hardcastle (1996) Optopalatography - A new apparatus for speech production analysis. Spoken Language ICSLP Proceedings. Vol. 3. 1589 - 1592.

Wrench, A.A. Electropalatography - Moving Forward. URL: http://epgresearch.com/info/assets/files/EPGSymposium.ppt/AlanWrenchppt.pdf, Last access: 30-09-2013.