# Automatic English phoneme recognition from articulatory data generated by EPG systems with grid and anatomical layout of contact sensors

Grzegorz Krynicki[1], Katarzyna Dziubalska-Kołaczyk[1], Jarosław Weckwerth[1], Grzegorz Michalski[1], Kamil Kaźmierski[1], Barbara Maciejewska[2], Bożena Wiskirska-Woźnica[2], Marzena Żygis[3], Wiesław Kuczko[4], Alicja Sekuła[2]

[1] Faculty of English, Adam Mickiewicz University in Poznan, Poland; [2] Department of Phoniatrics and Audiology, Poznan University of Medical Sciences; [3] Zentrum für Allgemeine Sprachwissenschaft Berlin, Germany; [4] Management and Production Engineering, Poznan University of Technology

## ABSTRACT

The aim of the study was to conduct automatic phoneme identification from articulatory data that accompanied the production of these phonemes in continuous speech. The articulatory data were obtained from 2 electropalatographic systems, *Palatometer* by *Complete Speech* and *Linguagraph* by *Rose-Medical*. *Palatometer* was used with the artificial palate containing 124 contact sensors in a grid layout, including 2 sensors monitoring the lip contact. The palate included a vacuum-thermoformed flexible printed circuit. *Linguagraph* was used with the acrylic artificial palate designed and developed for the purpose of this study, containing 62 electrodes in anatomical layout. *Palatometer* was used by one native of General American and *Linguagraph* by one native of General British, each reading 140 phonetically balanced sentences that included Harvard Sentences and TIMIT prompts. The EPG data were parametrised into dimensionality reduction indexes, which were analysed by means of linear discriminant analysis and a probabilistic neural network. The results of classifications are discussed.

**Keywords**: electropalatography, EPG, English, phoneme recognition, articulatory phonetics

## 1. INTRODUCTION

The concept of phoneme recognition refers to the identification of phonemes underlying an utterance without any reference to the language model at the word level or to a pronunciation dictionary.

Phoneme recognition from articulatory clues has mainly been used to improve predictions of missing information in the acoustic signal in robust automatic speech recognition and in the development of silent speech interfaces. Articulatory data have mainly been derived from one of three sources: acoustic-articulatory transformations using inverse mapping, classification scores for pseudo-articulatory features and direct physical measurements [10]. As the source of direct physical measurements, a wide range of techniques have been used, including X-ray filming [1], EGG, EMA and EPG [21] and EMA alone [7]. In all of the studies in which articulatory data were obtained from EPG, the palates with the contact sensor layout normalized for between-speaker anatomical differences were used.

In this paper, we report on one aspect of a larger project[5], whose aims include:

- testing experimentally whether teaching the pronunciation of English consonants and consonant clusters to Polish learners by means of direct visual feedback based on EPG information is more effective than a comparable method without EPG;
- conducting a systematic comparative analysis of differences between Polish and English with respect to articulatory correlates of distinctive features of the phonemes of the two languages;
- collecting an articulatory database that could be used to train robust automatic speech recognition systems and silent speech interfaces.

The last two aims required that we chose the type of the EPG system that would allow for the minimization of between-speaker articulatory variance in comparable pronunciations and that the selected system allows as reliable mapping of articulatory to acoustic distinctive features as it is possible considering the fact that the similar sounds can be created by a single speaker using a range of different articulatory gestures [15].

In the present report we use EPG articulatory data for phoneme recognition with the assumption that it provides an effective practical test of the quality of mapping of articulatory to acoustic distinctive features. We also compare the phoneme recognition rates obtained from two EPG systems and we describe the design and performance of a new artificial palate developed for the purpose of the project. The present paper is a continuation of the work presented in [11].

## 2. GRID AND ANATOMICAL LAYOUT OF CONTACT SENSORS

There major two approaches to laying out contact sensors on an EPG palate include a *grid layout* where

[5] The project and this study have been funded by the National Science Centre (grant no. 2013/11/B/HS2/03151)

each contact is positioned by a fixed distance from the neighbouring contacts and a *normalized anatomical layout* where the number of contacts does not vary and each contact is aligned with anatomical landmarks [23]. The two systems selected for this study were representative of both approaches:

- *Linguagraph* by *Rose-Medical Ltd*. [9] with the palate that has the normalized anatomical electrode layout, and
- *Palatometer* by *Complete Speech* [2] with the palate that has the grid layout of electrodes.

For the purpose of this study, the *Linguagraph* multiplexer was used with the normalized anatomical palate of a new design intended to make the palate thinner and more natural to speak with. The *Palatometer* multiplexer was used with the original *Complete Speech* grid layout palate.

The major advantage of the normalized anatomical layout is that a given contact sensor position for one speaker can be compared like-for-like with that of another speaker to a much greater degree than in the grid layout palate. In the grid layout, speaker palates that are wider, longer or more arched require more contacts to cover the whole palatal surface than in the normalized anatomical layout. This results in low between-speaker comparability of patterns that accompany the same pronunciations [22]. Phonetic distinctive features induced from grid layout articulatory data of one speaker would therefore be difficult to generalize to other speakers.

**Table 1**: Features of *Linguagraph*, *Palatometer* and Original palate

| | Linguagraph | Palatometer | Original |
|---|---|---|---|
| layout | anatomical | grid | anatomical |
| sensor position vs. anatomy | fixed | variable | fixed |
| sensor-to-sensor position | variable | fixed | variable |
| between-speaker comparability | high | low | high |
| average palate thickness on all electrodes | 1.5-2.5mm M=2.45 σ=.49 | ~.5mm | 1.7-3.4mm M=2.10 σ=.53 |
| materials used | acrylic | PET-G, polyimide | acrylic |
| durability | high | average | high |
| number of electrodes | 62 | 124 | 62 |
| production | manual | automatic | manual |

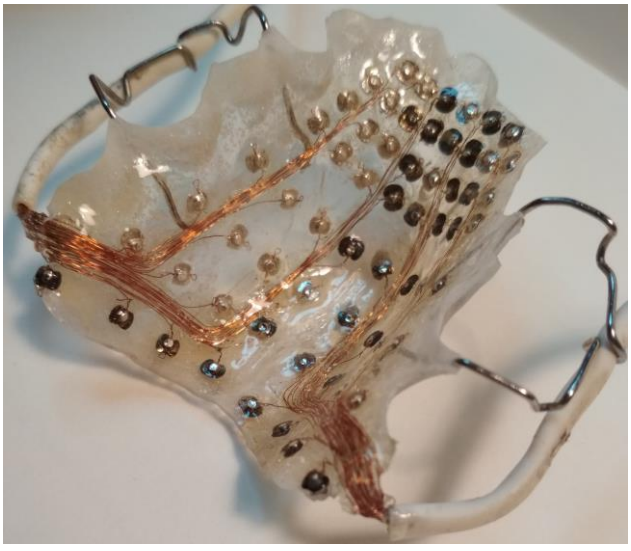
On the other hand, one of major disadvantages of normalized anatomical layout palates is that they are relatively thick. Thin grid layout palates increase the wearing comfort and result in a less distorted pronunciation. This, in turn, increases the chance that students using grid palates for learning L2 articulations would have a greater chance of preserving the correct articulation habits after the palate has been removed. The major disadvantage of the grid layout palates is, however, that the EPG patterns generated by L2 learners wearing them are difficult to reliably compare with native patterns.

For the purpose of this study, we have therefore developed a normalized anatomical palate with reduced thickness for low between-speaker variability without much compromising on the wearing comfort.

## 3. NEW TECHNIQUE OF MANUFACTURING NORMALIZED ANATOMICAL PALATES

The design mostly followed the traditional method of development of artificial palates with anatomical layout. The EPG system originally developed at Reading University in the mid 1980's [4] consists of 62 silver contacts embedded between two layers of two-component heat- and pressure-polymerized denture acrylic resin. Steel Adams clasps clip around the teeth retain the palate in place. Each of the contact sensors is soldered to a copper wire and the wires exit around the back of the rear molars. The two bundles of wires form each side of the palate are sealed in flexible tubing [22] and leave through the corners of the mouth.

**Figure 1**: Original palate with anatomical layout of contact sensors



In our approach (*Figure 1*), only one layer of conventional two-component acrylic is used. The

electrodes are covered with a thin layer of a one-component acrylic lacquer that is cured through UV radiation. This way we have reduced the mean thickness of the palate by approx. 0.35mm (*Table 1*, row 6, column 2 vs. 4; 1.5-2.5mm σ=.49 in the *Linguagraph* palate compared to 1.7-3.4mm with σ=.53). The palate thickness in all cases was measured as the average thickness at the location of all electrodes. We continue our work to reduce the relatively wide range of thickness values in our design. The technique of developing electrodes was also modified: instead of soldering a copper wire to a silver disc we inserted the copper wire into a silver bead, forged the bead into a bowler-shaped sensor and polished the crown to roughly level it with the surrounding acrylic. This method assured the firm hold of the electrode to the bottom layer of the acrylic by the brims.

## 4. APPROACHES TO PHONEME RECOGNITION

The state-of-the-art approaches to phoneme recognition include

- the hidden Markov model – Gaussian mixture modeling of phonemes [13] with additional discriminative training [3], and
- discriminative models e.g. recurrent neural networks [16], large margin classifiers [19] or multilayered perceptrons [18] have given higher phoneme recognition accuracies.

In this work, we estimate the posterior probabilities of phonemes based on articulatory information by means of two discriminative models, forward-selection linear discriminant analysis (LDA) and a four-layer probabilistic neural network (PNN). For classification by means of LDA, *Statistica* statistical package was used. For classification by means of PNN, we used *Statgraphics*.

## 5. ARTICULATORY DATA

For comparison with previous studies [19, 21] and to avoid problems related to between-speaker variability, the data in this study were obtained from one speaker per EPG system and covered about 34 minutes of speech in total. *Palatometer* was used by a 24-year-old female speaker of General American. *Linguagraph* with the original palate was used by a 41-year-old male speaker of General British. Both speakers read a list of 197 items: 10 sentences from the List 11 of Harvard Sentences [6], 130 phonetically balanced sentences from the TIMIT prompt list [20], English alphabet and numerals from 0 to 30. The prompts were manually adjusted to provide for minor fillers, re-starts and minor misreadings on the part of the speakers. The whole list in each case contained approx.. 1250 words. The database for General American contained 16 min. of audio data and 18 minutes for General British. The recordings were then segmented and annotated with phoneme labels using *Penn Forced Aligner* [8] for General American and for General British – *FAVE-align* [17] with the adapted British English dictionary *BEEP* [14] and the General American language model. The annotation was subjected to minor manual corrections.

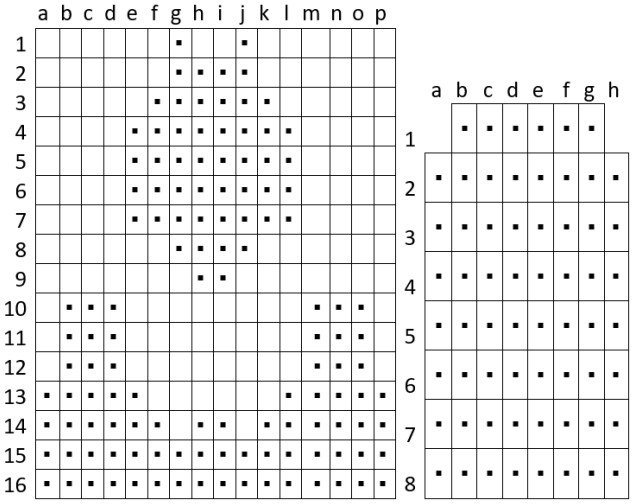
## 6. PARAMETRISATION OF THE EPG DATA

The EPG information was transformed into a set of linguistically meaningful and computationally manageable parameters – dimensionality reduction indices (DRI) adapted from Hardcastle et al. [5]. DRI's for a given phoneme were calculated for the single EPG frame at mid time of the phoneme duration. The DRI's were:

- *bilabial region* – sum of contacts in the 1st row of *Palatometer* (abbr. P below) pattern divided by all electrodes in that row; the 1st row of the P palate is labelled as 1 in the left part of *Figure 2* below; absent from *Linguagraph* – abbr. as L below;
- *dental region* – sum of contacts in rows P2-4, divided by all electrodes in these rows; absent from *Linguagraph*;
- *alveolar to prepalatal* – a sum of contacts in rows P5-9/L1-5 divided by all electrodes in these rows;
- *midpalatal to velar* – sum of contacts in rows P10-15/L6-8 divided by all electrodes in these rows;
- *total percentage of contacts* – sum of contacts in rows P1-15/L1-8 divided by all electrodes in these rows;
- *general centre of gravity* i.e. the weighted average of the sum of contacts in rows P1-15, where the weights on rows are 14, 13, ..., 1 respectively, and the weighted average of the sum of contacts in rows L1-8 where the weights on rows are 7, 6, …, 1 respectively;
- *anterior centre of gravity* i.e. the weighted average of the sum of contacts in rows P2-8 columns G-J where the weights are 14, 13, ..., 8 respectively, and the weighted average of the sum of contacts in rows L1-4 for columns C-F with weights 7, 6, 5 and 4 respectively;
- *posterior centre of gravity*, i.e. the weighted average of the sum of contacts in rows P9-15 columns D-M with the weights on rows 7, 6, ..., 1 respectively, the weighted average of the sum of contacts in rows L4-7 columns B-G, with weights on rows 4, 3, 2 and 1 respectively;

- *laterality* i.e. the weighted average of the sum of contacts in rows P1-15 columns D-M, where the weights on columns are 1, 2, ..., 5 in columns D-H and 5, 4, ...,1 in columns I-M. In rows L1-8 columns B-G, where the weights on columns are 1, 2 and 3 in columns B-D and weights 3, 2 and 1 in columns E-G respectively;
- *asymmetry* i.e. the difference between the sum of contacts in columns A-H and I-P for *Palatometer* and the difference between the sum of contacts in columns A-D and E-H for *Linguagraph*;
- *fricativity* i.e. the sum of contacts in rows P2-5 divided by all electrodes in these rows minus the sum of contacts in rows P6-8 divided by all electrodes in these rows for *Palatometer* and the sum of contacts in rows L1-2 divided by all electrodes in these rows minus the sum of contacts in rows L3-4 divided by all electrodes in these rows for *Linguagraph.*

**Figure 2**: *Palatometer* (left) and *Linguagraph* electrode layout; dots indicate electrode locations. Electrodes in *Palatometer* row 10, and col. a and p for rows 13-14 were never activated.




## 7. CLASSIFICATION PROCEDURE AND RESULTS

For *Palatometer* – 3210 unique phoneme-DRI pairs were used to develop classification models that discriminated among 38 phonemes. As in [21], consonants were classified along with vowels. For *Linguagraph*, 3287 unique phoneme-DRI pairs were used to develop classification models that discriminated among 43 phonemes. In both cases, the /ʒ/ phoneme was excluded from all classifications as it was illustrated by only 2 cases. In the case of PNN classification, jack-knifing (leave-one-out) method was used as a cross-validation technique. In PNN, prior probabilities used were proportional to the observed. The table with results of the classification of individual phonemes can be accessed at [12]. The results of the general classification is presented in *Table 2*. The best classifier for any EPG system used in the study was PNN and it performed with 32.1% correct classification rate for the data obtained with *Palatometer*.

**Table 2**: Classification results for DRI's calculated for articulatory data from two EPG systems

| EPG system | PNN | LDA |
|---|---|---|
| Linguagraph + original anatomical palate | 30.8 | 29.6 |
| Palatometer | 32.1 | 31.3 |

The data obtained by *Palatometer* were generally classified at marginally better rate than the data obtained by *Linguagraph*.

## 8. CONCLUSIONS

The higher results of phoneme classification based on the data obtained from *Palatometer* in comparison to *Linguagraph* may be attributed to the fact that *Palatometer* has a higher density of the contact sensors, i.e. 124 electrodes (114 activated at least once) vs. 62 in *Linguagraph*. Moreover, *Palatometer* has 2 bilabial and 10 dental sensors that palate designed to work with *Linguagraph* did not have. Finally, the accuracy of forced alignment for American data may be higher than that for British data.

The correct classification rates presented in *Table 2* are relatively low compared to the results reported in [21]: phone error rate of 35.7% for monophone recognition on TIMIT data using all sources of physical measurement, i.e. EMA, EGG and EPG. It should however be noted that in [21] EPG data yields only 1.5% absolute PER reduction to the result obtained based on EMA and EGG and that authors do not report PER based on EPG alone. Moreover, in [21], 30 minutes of speech was used compared to 16-18 minutes used in the present study. Our results may thus suggest that:

- the classification of 38-43 phonemes based on 3210-3387 phoneme-DRI pairs is challenging. Multiple coarticulatory and random effects cause DRI's to vary greatly. Preliminary experiments on General British data [12] indicate that the results of PNN classification reach 49.46% if trained on a set of 24985 phoneme-DRI pairs obtained with *Linguagraph* and our original anatomical palate;
- the phoneme classification problem is difficult to solved without considering the probabilities of phoneme or word sequences (e.g. in the form of

Hidden Markov models and N-gram models, c.f. [21]);

- in further tests, additional methods of deriving DRI's should be considered, e.g. principal component analysis;
- recognition could be further improved by providing additional sources of physical measurement.

Still, the results show that the tested EPG systems, irrespective of whether their electrodes are arranged in the anatomical or grid layout, both can be used for phoneme recognition, and consequently, both allow comparably reliable mapping of articulatory to acoustic distinctive features. We have also demonstrated that, with respect to the phoneme classification results, the original anatomical palate developed for the purpose of the project the present study is a part of, is on a par with the commercially available solution.

## 9. REFERENCES


[1] Blackburn, C and Young, SJ. 2001. Enhanced speech recognition using an articulatory production model trained on X-ray data. Computer Speech and Language, 15. pp. 195-216.

[2] Complete Speech. 2018. The SmartPalate System. Retrieved from https://completespeech.com Last access: 05-12-2018.

[3] Fu, Q., X. He, and L. Deng. 2007. Phone-Discriminating Minimum Classification Error (P-MCE) Training for Phonetic Recognition. Proc. of Interspeech.

[4] Hardcastle, W. J., Gibbon, F. E., & Jones W. 1991a. Visual display of tongue-palate contact: Electropalatography in the assessment and remediation of speech disorders. British Journal of Disorders of Communication, 26, 41 – 74.

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

[6] Harvard Sentences. 2018. URL: http://www.cs.columbia.edu/~hgs/audio/harvard.html Last access: 05-12-2018.

[7] Heracleous, Panikos, Pierre Badin, Gérard Bailly, Norihiro Hagita. 2011. A pilot study on augmented speech communication based on Electro-Magnetic Articulography. Pattern Recognition Letters 32(8): 1119-1125.

[8] Jiahong, Y. - M. Liberman. 2008. Speaker identification on the SCOTUS corpus. Proceedings of Acoustics 2008: 5687-5690.

[9] Kelly, S. – A. Main – G. Manley – C. McLean. 2000. Electropalatography and the Linguagraph system. Medical Engineering & Physics Vol. 22/1. Str. 47-58.

[10] Kirchhoff, K., G.A. Fink, G. Sagerer. 2002. Combining acoustic and articulatory feature information for robust speech recognition. Speech Commun. 37, 303–319.

[11] Krynicki, Grzegorz. 2014. Articulatory grounding of phonemic distinctions in English by means of electropalatography. In: Eugeniusz Cyran and Jolanta Szpyra-Kozłowska (Eds) Crossing Phonetics-Phonology Lines. Newcastle Upon Tyne: Cambridge Scholars Publishing. Pp. 299-312.

[12] Krynicki, Grzegorz. 2019. LDA classification table and pair-wise phoneme classifications, URL: http://wa.amu.edu.pl/~krynicki/pub/icphs2019

[13] Lee, K.-F. and H.-W. Hon. 1989. Speaker-Independent Phone Recognition using Hidden Markov Models. IEEE Trans. Acoust. Speech Signal Process., vol. 37, no. 11, pp. 1641–1648.

[14] MacKenzie, Laurel. 2016. Personal communication

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

[16] Robinson, A. 1994. "An Application of Recurrent Nets to Phone Probability Estimation. IEEE Trans. on Neural Networks, vol. 5, no. 2, pp. 298–305.

[17] Rosenfelder, Ingrid; Fruehwald, Josef; Evanini, Keelan; Seyfarth, Scott; Gorman, Kyle; Prichard, Hilary; Yuan, Jiahong; 2014. FAVE (Forced Alignment and Vowel Extraction) Program Suite v1.2.2

[18] Schwarz, P., P. Matejka and J. Cernocky. 2006. Hierarchical Structures of Neural Networks for Phoneme Recognition. Proc. of IEEE Conf. Acoust. Speech. Signal Process. (ICASSP).

[19] Sha, F., and L. Saul. 2006. Large Margin Gaussian Mixture Modeling for Phonetic Classification and Recognition. Proc. of IEEE Conf. Acoust. Speech. Signal Process. (ICASSP).

[20] TIMIT. 2018. URL: https://catalog.ldc.upenn.edu/docs/LDC2008S09/manual/html/node16.html, Last access: 05-12-2018.

[21] Uraga, E. and Hain, T., 2006. Automatic speech recognition experiments with articulatory data. Proc. INTERSPEECH, 353-356.

[22] Wrench A. 2001. A new resource for speech production modelling in speech technology. Proceedings of the Workshop on Innovation in Speech Processing.

[23] Wrench, Alan A. 2007. Advances in EPG palate design. In: Advances in Speech-Language Pathology, 9(1). Pp. 3-12.