Available online at www.sciencedirect.com
ScienceDirect

Computer Speech and Language 22 (2008) 346–359

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Estimating electropalatographic patterns from the speech signal

Asterios Toutios [a,*,1], Konstantinos Margaritis [b]

[a] Equipe Parole, LORIA, Campus Scientifique – BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France
[b] Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece

Received 14 July 2006; received in revised form 4 December 2007; accepted 5 December 2007
Available online 23 December 2007

Abstract

Electropalatography is a well established technique for recording information on the patterns of contact between the tongue and the hard palate during speech, leading to a stream of binary vectors representing contacts or non-contacts between the tongue and certain positions on the hard palate. A data-driven approach to mapping the speech signal onto electropalatographic information is presented. Principal component analysis is used to model the spatial structure of the electropalatographic data and support vector regression is used to map acoustic parameters onto projections of the electropalatographic data on the principal components.

Keywords: Electropalatography; Speech inversion; Support vector regression; Principal component analysis

1. Introduction

Electropalatography (EPG) (Gibbon and Nicolaidis, 1999) is a widely used technique for recording and analyzing one aspect of tongue activity, namely its contact with the hard palate during continuous speech. It is well established as a relatively non-invasive, conceptually simple and easy-to-use tool for the investigation of lingual activity in both normal and pathological speech. An essential component of EPG is a custom-made artificial palate, which is molded to fit as unobtrusively as possible against a speaker's hard palate. Embedded in it are a number of electrodes: 32, 62, 64, 96 or 128, depending on the implementation (Hiiemae and Palmer, 2003). When contact occurs between the tongue surface and any of the electrodes, a signal is conducted to an external processing unit and recorded. Typically, the sampling rate of such a system is 100–200 Hz. Thus, for a given utterance, the sequence of raw EPG data consists of a stream of binary vectors with both spatial and temporal structure. Usually a value of 1 represents the event that the tongue contacts a given electrode at a given point in time and a 0 value that it does not. (However, in this paper, such a non-contact will be

[*] Corresponding author. Tel.: +33 383 59 30 32; fax: +33 383 55 25 73.
E-mail addresses: toutiosa@loria.fr (A. Toutios), kmarg@uom.gr (K. Margaritis).

[1] The work was supported by the Greek research program "HRAKLEITOS", which is co-funded by the European Social Fund (80%) and National Resources (20%).

0885-2308/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2007.12.002

represented by a value of −1.) Observation of both temporal and spatial details of contact across the entire palatal region can be very helpful to identify many phonetically relevant details of lingual activity. EPG has been successfully used to study a number of phenomena in phonetic descriptive work, in studies of lingual coarticulation and in the diagnosis and treatment of a variety of speech disorders (Gibbon, 2005).

The freely available MOCHA database (Wrench and Hardcastle, 2000) is a source of articulatory data recorded in parallel with the corresponding acoustic information. It includes four data streams recorded concurrently: the acoustic waveform, sampled at 16 kHz with 16 bit precision, laryngograph, electromagnetic articulograph and electropalatograph data. For the latter, the reading EPG (Jones and Hardcastle, 1995) system is used, which provides tongue-palate contact data at 62 normalized positions on the hard palate, defined by landmarks on the maxilla. These are recorded at 200 Hz. Speakers read a set of 460 British TIMIT-style (Garofolo et al., 1993) sentences, which are designed to provide phonetically diverse material and capture the connected speech processes in English with good coverage. All waveforms are labeled at the phonemic level; however, these labels are the result of a forced-alignment process and considered prone to errors (according to the documentation of the database, 6% of the labels are wrong). The original plan was that the database would feature up to 40 speakers with a variety of regional accents. Up to the time of writing this paper data from ten speakers are available. Data from three of these speakers are also fully corrected.

This paper presents a machine learning-based method for estimating electropalatographic information from the acoustic waveform using data from the MOCHA database. The method is data-driven, in the sense that no a priori expert speech production knowledge is employed. At testing, the trained system does not need to explicitly know the phonetic identity of the sound it processes in order to estimate EPG patterns from the corresponding acoustic information.

The method involves three steps. Firstly, the application of principal component analysis to the EPG data. Secondly, mapping acoustic parameters to the projections of the EPG data on the principal components using support vector regression. Thirdly, the production of estimates of the EPG sequences from the estimates of the projections on the principal components.

The task of mapping the speech signal to EPG patterns, may be considered as a special case of the acoustic-to-articulatory mapping problem, or speech inversion (Schroeter and Sondhi, 1994), which refers to estimating articulatory parameters based on the corresponding acoustic information, with a possible addition of complementary cues like measures taken by video captures of the speaker's face (Engwall, 2005). The problem has received considerable attention in the speech community, having numerous possible applications in the areas of speech coding, speech recognition, speech therapy, language acquisition and second language learning (Richmond et al., 2003; Ouni and Laprie, 2005), as well as increasing understanding of the relations between speech production and acoustics from a phonetic viewpoint (Mokhtari et al., 2007). What we imply is that a successful mapping from acoustics to EPG information may have all the applications attributed to the general acoustic-to-articulatory mapping. For example, a successful acoustic-to-EPG mapping might allow the remediation of numerous speech disorders by visual feedbacks of tongue activity (Gibbon, 2005) in cases where the cost of an artificial palate is prohibitive. Another application might be the analysis of the tongue-palate contact patterns in the recorded speech of people who are unable or unwilling to wear an artificial palate.

The rest of the paper is organized as follows. Firstly, we discuss principal component analysis and support vector regression, and describe the processing steps applied to our data. We then elaborate on our method and present experimental results, concluding with a series of observations on the results and their implications.

2. Principal component analysis

Principal component analysis (PCA) (Jolliffe, 1986) is a well-known statistical method which projects some data onto a new set of axes. These axes are the directions in the data space where the data variation is maximum and they are called the *principal components*. The principal components are ordered by significance. Dimensionality reduction of the data may be achieved by eliminating the least significant principal components. Practically, PCA is accomplished by applying eigenvalue analysis on the covariance matrix of the data. The eigenvectors are then the principal components.

Even though PCA is not an explicitly binary method, it has been used to model the spatial structure of EPG data in the past (Nguyen et al., 1996; Carreira-Perpiñán and Renals, 1998). In both the cases, PCA modeling is considered successful; however, not the optimal method among the ones studied. In particular, Nguyen et al. (1996) come to propose an autoassociative neural network-based dimensionality reduction scheme and Carreira-Perpiñán and Renals (1998) the Generative Topographic Mapping. Regarding the number of principal components used to model the data, Nguyen et al. (1996) present results for eight principal components and Carreira-Perpiñán and Renals (1998) depict nine principal components. Nevertheless, the actual intrinsic dimensionality of the EPG data is a matter under investigation, with a dimension of 5–10 usually suggested (Carreira-Perpiñán and Renals, 1998).

PCA has several useful properties including computational simplicity, the additivity of the principal components, their straightforward visualization and the ease of resynthesizing the original EPG patterns from the corresponding projections. These constitute the main reason for choosing to utilize PCA for our purposes, among all other possible spatial modeling schemes.

3. Support vector regression

Support vector regression (SVR) (Smola and Schölkhopf, 2004) is a supervised regression learning method that has been shown to produce state-of-the-art results for several regression problems; including a study of ours on mapping the speech signal onto electromagnetic articulograph information (Toutios, 2006). Being non-linear, it is generally accepted as a very powerful alternative to neural networks and other regression algorithms. There are several instances of the method, the original one being the ε-SVR algorithm (Vapnik, 1995).

Given n training vectors $\mathbf{x}_i$ and real-valued corresponding outputs $y_i \in R$, one wants to find an estimate for the function $y = f(\mathbf{x})$. According to ε-SVR, this estimate is

$$f(\mathbf{x}) = \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{a}_\mathbf{i}^* - \mathbf{a}_\mathbf{i})\mathbf{k}(\mathbf{x}_\mathbf{i}, \mathbf{x}) + \mathbf{b}, \tag{1}$$

where the coefficients (Lagrange multipliers) a_i and a_i^* are the solution for the quadratic optimization problem

$$\begin{gathered}\text{maximize}\\ W(\mathbf{a}, \mathbf{a}^*) = -\varepsilon \sum_{i=1}^{n} (a_i^* + a_i) + \sum_{i=1}^{n} (a_i^* - a_i) y_i - \frac{1}{2} \sum_{i,j=1}^{n} (a_i^* - a_i)(a_j^* - a_j) k(\mathbf{x}_i \mathbf{x}_j)\\ \text{subject to}\\ 0 \leqslant a_i,\ a_i^* \leqslant C,\ i = 1, \ldots, n,\ \text{and}\ \sum_{i=1}^{n} (a_i^* - a_i) = 0.\end{gathered} \tag{2}$$

where $C > 0$ and $\varepsilon > 0$ are pre-selected constants.

The kernel function $k(.,.)$ maps input data into a higher-dimensional space to account for non-linearities in the function to be estimated (1). The usual choice for the kernel function is the Radial Basis Function (RBF) kernel

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma^2}\right), \tag{3}$$

with the parameter σ being a pre-selected constant. Other choices for the kernel function include linear, polynomial and hyperbolic tangent kernels.

In this paper, SVR training is carried out with the SVMTorch II software (Collobert and Bengio, 2001). The RBF kernel is used throughout. The parameters, C, ε and the kernel parameter σ, are estimated from the training dataset using the following heuristics, which are loosely based on Cherkassky and Ma (2004), Weston et al. (2003) and our own experience. For the parameter C, we use

$$C = \max(|\ \bar{y} + 3\sigma_y\ |, |\ \bar{y} - 3\sigma_y\ |), \tag{4}$$

where $\bar{y}$ and σ_y are the mean and the standard deviation of the output values of training data, respectively. For the parameter ε, we use

$$\varepsilon = 3\sigma_n\sqrt{\frac{\ln n}{n}}, \tag{5}$$

where n is the number of training examples, and σ_n is the median value of $\sqrt{(y-\bar{y})^2}$ across the training output data. Finally, for the RBF kernel parameter σ, we use

$$\sigma = \sum_{j=1}^{d} \sigma_{x_j}, \tag{6}$$

where σ_{x_j} is the standard deviation of the jth element of the input data vectors across the training set.

4. Data processing

The MOCHA database includes 460 utterances spoken by the fsew0 speaker, a female with a Southern English accent. The following processing steps are applied.

Firstly, based on the MOCHA label files, silent parts from the beginning and end of the utterances are omitted. Secondly, the EPG data are undersampled from 200 Hz to 100 Hz. Next, using the HTK Toolkit (Young et al., 2005), 12 MF-PLPs (Woodland et al., 1997) and log energy are extracted from the speech signal, using 16 ms windows with 10 ms shifts and 40 filterbanks. These acoustic parameters are centered around their mean and scaled by their standard deviation. The result of this process is 124, 242 pairs of 62-dimensional EPG vectors and 13-dimensional acoustic vectors.

From the 460 available utterances which are numbered and presented in the list provided at http://data.cstr.ed.ac.uk/mocha/mocha-timit.txt, 92 (every 10th utterance beginning with the 2nd and every 10th one beginning with the 6th, 24,388 acoustic-EPG pairs) are reserved for testing. The rest (99,904 acoustic-EPG pairs) constitute the training dataset Fig. 1.

5. Application of PCA on EPG data

Principal component analysis is applied on the EPG data of the training dataset. Fig. 2a is the scree plot of this analysis, that is, a plot of the eigenvalues versus the order of the associated principal component (up to the 62nd), as an indication of the proportion of the variance of the EPG data explained by each principal component.

At the PCA reconstruction phase, calculated EPG patterns are converted into binary vectors by considering the signum of their 62 elements, which are originally continuous. Then a reconstruction error index is defined by computing the number of electrodes for which the reconstructed value is wrong – that is, a positive (contact) instead of a negative (non-contact) or vice-versa – in each reconstructed EPG vector. We measure the mean reconstruction error when the EPG data are reconstructed from the projections on the L ($1 \leqslant L \leqslant 62$), the most significant principal components. The mean reconstruction error on the test dataset

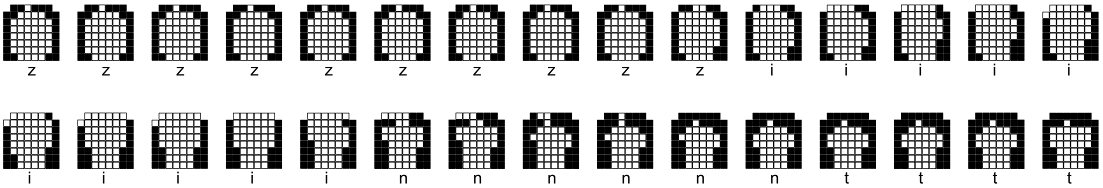

Fig. 1. Part of typical EPG sequence. The shape of the figures (EPG vectors or electropalatograms) follows that of the palate, the alveolar part being at the top and the velar part at the bottom. Black squares indicate a contact between the tongue and the palate. This is from the utterance "The hallway opens into a huge chamber", EPG sequence corresponding to the part in boldface. The speaker is fsew0 from the MOCHA database. Corresponding MOCHA labels are shown.

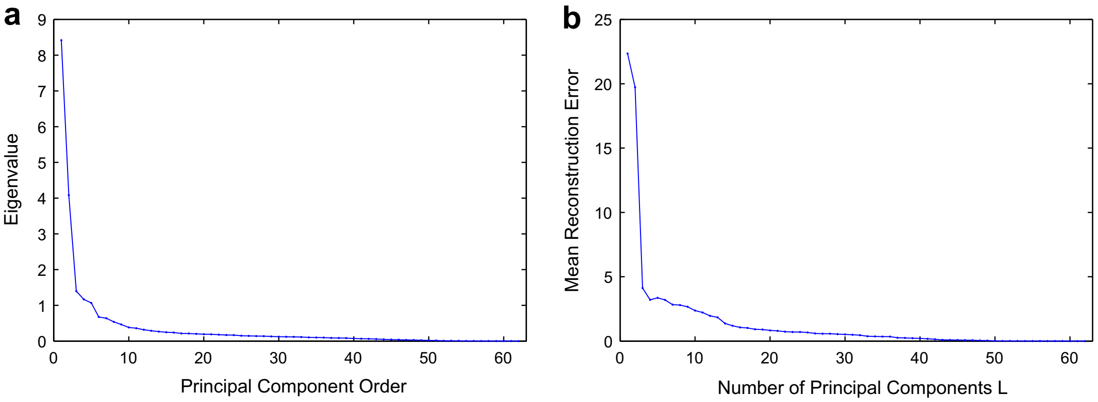

Fig. 2. (a) Scree plot of principal component analysis of EPG data; and (b) reconstruction error on the test data when the projections on the *L* most significant principal components are used, calculated as mean number of “misclassified” EPG elements.

is shown in Fig. 2b. We note that the corresponding curve (not shown) for the training dataset is almost identical.

For the rest of this paper we will regard the projections on the first nine principal components, which lead to adequately small reconstruction error, and in accordance with the previously mentioned works on latent modeling of EPG data (we will further comment on this choice later). Fig. 3 shows these principal components. Fig. 4 provides indications of the distributions of the values of the projections of the EPG data on the first eight principal components, according to the different phonemic labels as labeled in the MOCHA database. The correspondence of these labels to IPA symbols is shown in Table 1. Fig. 5 shows rough schematics of the “mean” EPG patterns across the test set for each phonemic label.

Examination of Fig. 4 may partly reveal the systematic features captured by the principal components. For example, the top left plot (PC1 vs. PC2) shows that alveolars (t, d, n, s, z), affricates (ch, jh) and postalveolars (sh, zh) are quite distinguishable from the rest of the sounds, defining a sub-area of their own in the graph characterized by small values on both these principal components. The dentals (th, dh) also seem to define another particular sub-area, with larger values on 1st principal component. In the top-right plot (PC3 vs.

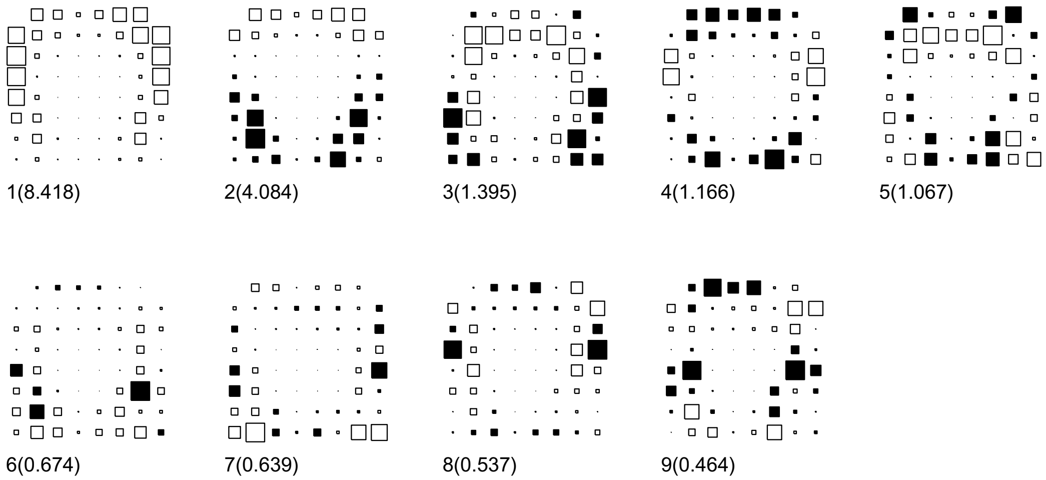

Fig. 3. First nine principal components of the EPG data. Each value is represented by a square of size proportional to its absolute value and color black or white whether it is positive or negative. Numbers in parentheses are the corresponding eigenvalues.

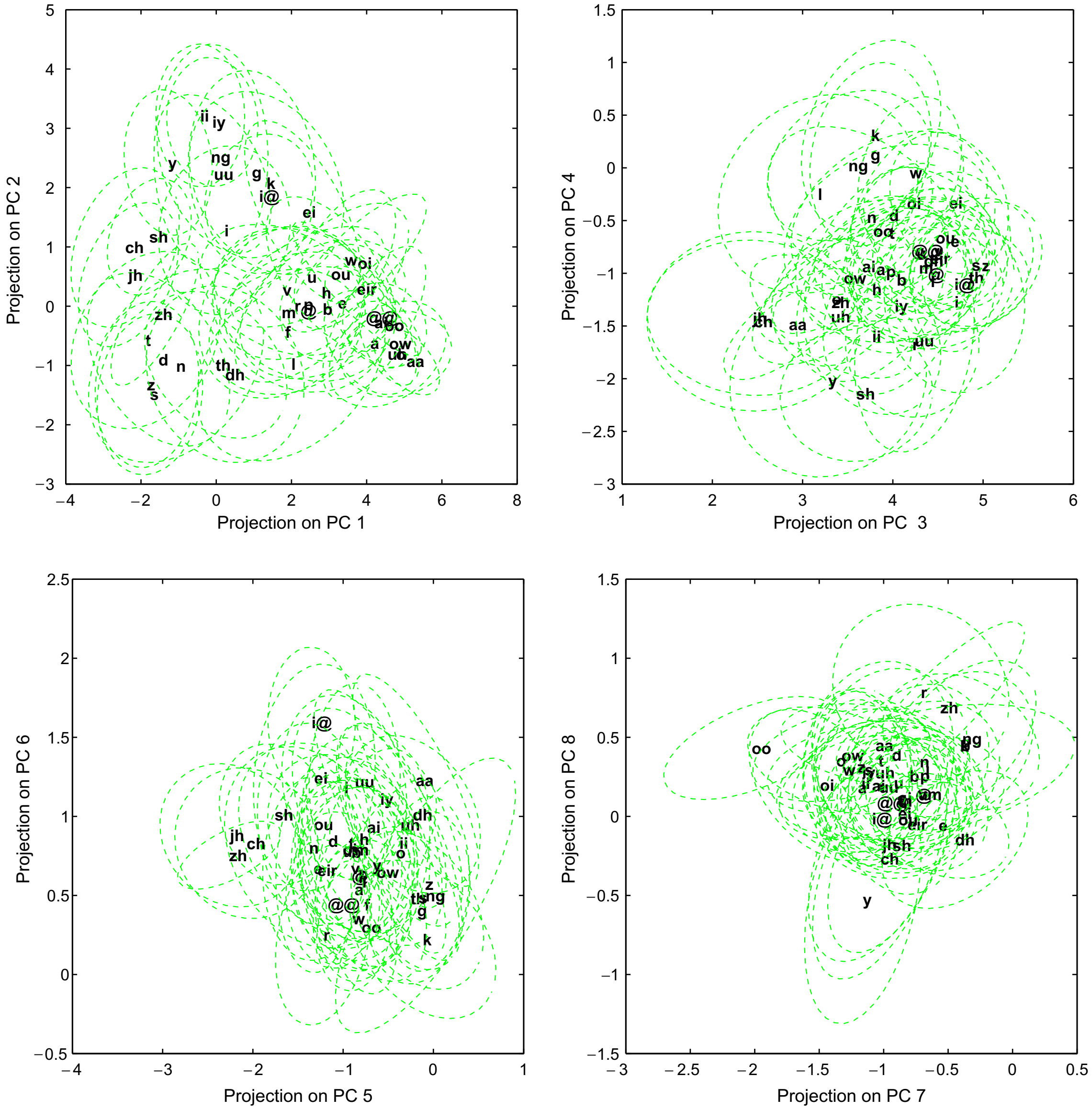

Fig. 4. Dispersions of projections on principal components according to phonemic labels (test set). MOCHA labels are located at the point of the mean value of the projection across instances of the labels.

PC4), the velars (k, g, ng) are pointed out, mostly attributable to the 4th principal component. In the bottom-left plot (PC5 vs. PC6) the affricates and postalveolars are adequately distinct from the rest of the sounds. Considering vowels, Fig. 5 indicates a difference between close vowels (i, ii, iy, uu) and the rest of vowels, with the former showing great degree of lateral contact. This difference may also be observed in Fig. 4, especially in the PC1 vs. PC2 plot.

Regarding the efficiency in reconstructing EPG patterns from the principal components, Fig. 6 plots the reconstruction error against the total number of contacts in each EPG pattern, again broken down by phonemic label. The general rule of thumb is that the more contacts in a pattern, the bigger the error in reconstructing it from the principal components. Or, phonemes exhibiting large total number of contacts are most difficult to model by PCA.

Table 1
Phonemic labels in the MOCHA database and their corresponding IPA symbols

Front vowels	a(æ)	e(ɛ)	i(ɪ)	ii(iː)	iy(i)		
Mid vowels	@(ə)	@@(ɚ)	uh(ʌ)				
Back vowels	aa(ɑ)	o(ɒ)	oo(ɔ)	u(ʊ)	uu(u)		
Diphthongs	ai(aɪ)	ei(eɪ)	eir(ɛə)	i@(ɪə)	oi(ʊɪ)	ou(oʊ)	ow(aʊ)
Bilabial	p(p)	b(b)	m(m)				
Labiodental	f(f)	v(v)					
Dental	th(θ)	dh(ð)					
Alveolar	t(t)	d(d)	n(n)	s(s)	z(z)		
Alveolar (affricates)	ch(tʃ)	jh(dʒ)					
Alveolar (approximants)	l(l)	r(ɹ)					
Postalveolar	sh(ʃ)	zh(ʒ)					
Velar	g(g)	k(k)	ng(ŋ)				
Glottal	h(h)						
Glides	w(w)	y(j)					

Partly adapted from Frankel (2003).

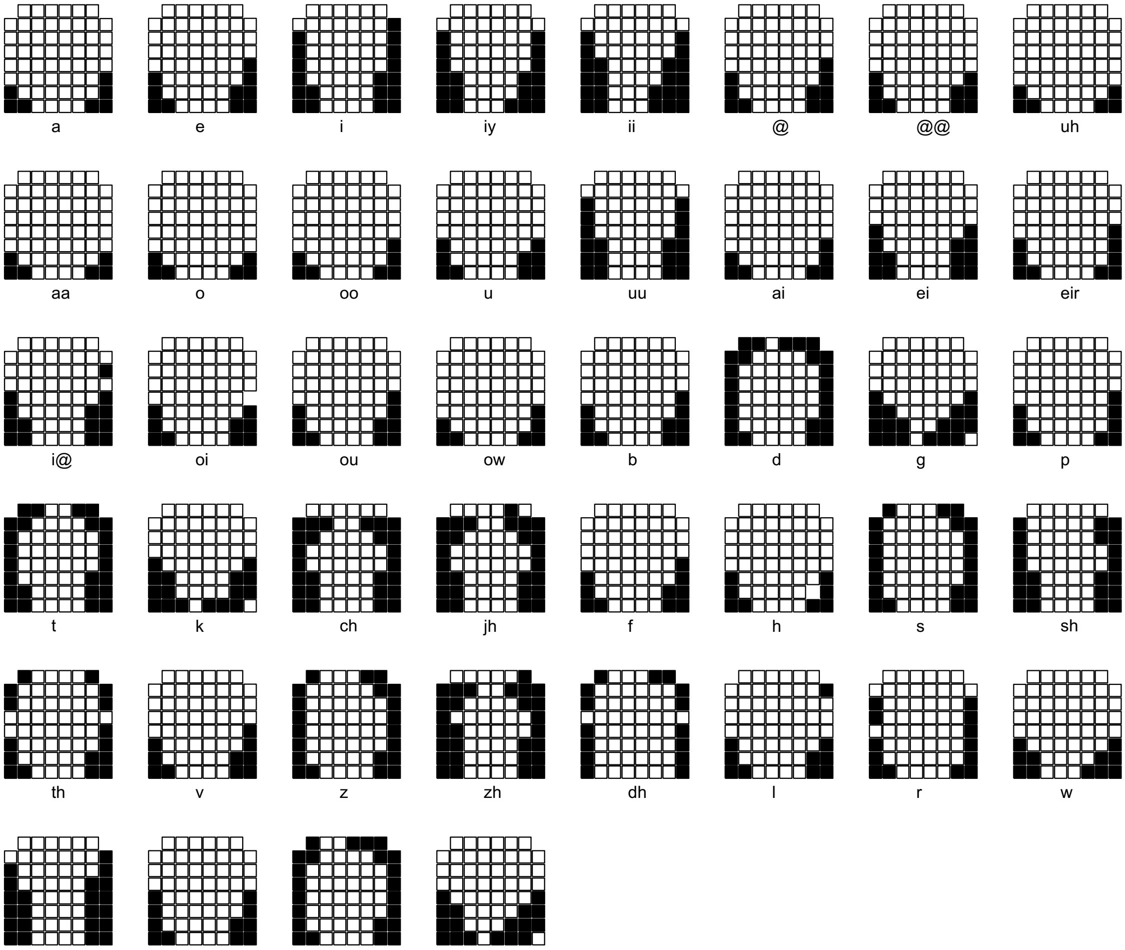

Fig. 5. "Mean" EPG patterns across MOCHA labels in the test set. An electrode is black when it is a contact for the majority of patterns with the specific label.

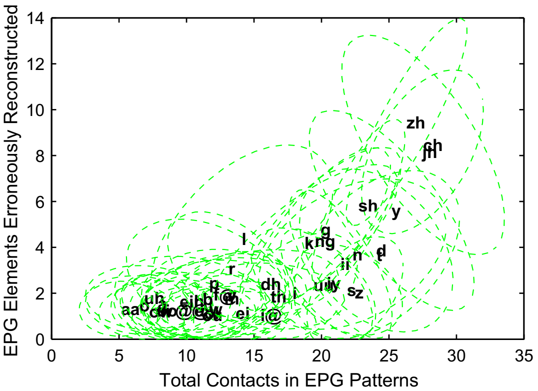

Fig. 6. PCA reconstruction error plotted against total number of contacts in EPG patterns and broken down by phonemic label. MOCHA labels are located at the point of the mean value of number of total contacts and reconstruction error across instances of the labels.

6. Estimating projections from acoustic parameters

While the principal components are constant throughout the dataset, the projections of the EPG data on the principal components are functions of time. Our goal from this point on is to produce estimates of the latter projections from the speech signal. To this end we use the ε-SVR algorithm, as described, considering the projections on one principal component at a time.

As other works on acoustic-to-articulatory mappings suggest (Richmond et al., 2003), the input vectors of the regression algorithm should include acoustic information spanning over a relatively large context window. A first question to answer is how large this window should be for our task. We train estimate regression functions for the projections on the first nine principal using incremental sizes of context windows. We consider only symmetric context windows, that is, the same number of context frames (from 0 up to 10) is incrementally added before and after the acoustic frame exactly corresponding to the output value. For this experiment, only a small fraction of the training dataset (roughly the first out of 10 examples in sequence) is actually used for training. The estimated functions are tested against the whole test set. The normalized mean squared error is measured

$$\text{NMSE} = \frac{1}{\sigma_y}\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i' - y_i)^2}, \tag{7}$$

where y and y' are actual and estimated projections, respectively, m is the number of test examples and σ_y is the standard deviation of y.

The NMSE for the projections on the first nine principal components as a function of the number of context frames added before and after the frame in question is shown in Fig. 7, as well as its mean across the nine projections. The minimum mean value is achieved with seven frames added before and after. That is, context windows that minimize NMSE consist of 15 acoustic frames (frame in question + 7 frames before + 7 frames after) and span over roughly 161 ms of speech.

Using these context windows for constructing input vectors, one support vector estimation function for each of the nine projections is trained on the complete training dataset. Fig. 8 shows the NMSE and the Pearson product-moment correlation between actual and estimated projections as functions of the corresponding principal components. Especially for the 1st and 2nd component, the results are very good, demonstrating that the corresponding projections are estimated quite efficiently from the speech signal. Results for components 3–5 might also be considered satisfactory, while estimation of projections on the principal components 6–9 seems relatively poor. Fig. 9 shows actual and estimated projections on the principal components for one utterance from the test set, where fsew0 utters the phrase “The hallway opens into a huge chamber”. The corresponding MOCHA labels are also included.

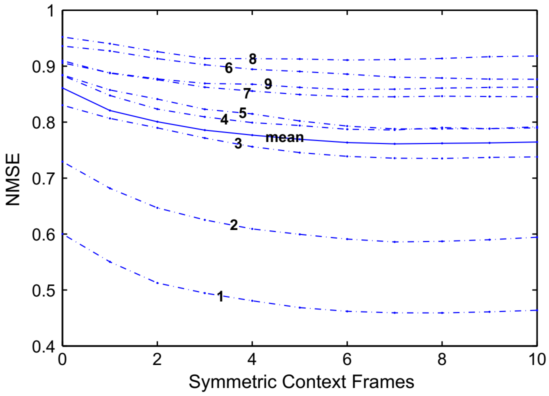

Fig. 7. Results of a small experiment (using roughly the first out of 10 examples in sequence from training set) for finding the size of the optimal context window for the mapping from acoustic parameters to the first nine principal components. Normalized mean squared error versus number of symmetric context frames added before and after the frame in question. The dash-dotted curves correspond to the first nine principal components, as numbered. Solid curve is the mean value across principal components. Minimum is achieved with seven frames.

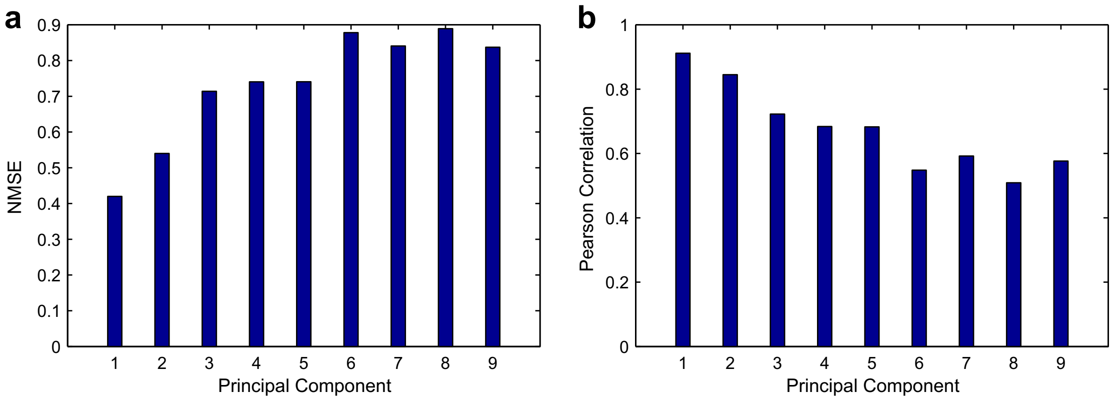

Fig. 8. Normalized mean squared error and Pearson product moment correlation between actual and estimated projections as functions of the corresponding principal components.

7. From estimated projections to estimated EPG sequences

For the final phase of our method, the EPG patterns are reconstructed from the *estimated* projections on the principal components. These *estimated* EPG patterns are again converted into binary vectors by considering the signum of their elements. We stress that these estimated patterns are essentially produced from the acoustic parameters. An estimation error index is defined in the same manner as the previous reconstruction error index using this time the new estimated EPG vectors instead of the PCA reconstructed ones. Fig. 10a presents this estimation error as a function of the number of projections on the principal components used to estimate the EPG patterns. Beginning with a value of 22.75 using only the 1st component (see also Table 2), the error decreases at 20.77 when the 2nd component is added, followed by a dramatic drop at 5.22 with the 3rd component. The error decreases a little more with the 4th component, slightly increases with the 5th component, and then decreases until it reaches a plateau after the 7th component. In Fig. 10b, we plot the mean difference in the EPG patterns derived using L components compared to those derived using $L-1$ components. This shows that, in this setup, the inclusion of further principal components alters the estimated EPG patterns more significantly than Fig. 10a suggests. Table 2 summarizes, in numerical form, the results already presented in Fig. 2 (for the first nine principal components), Figs. 8 and 10.

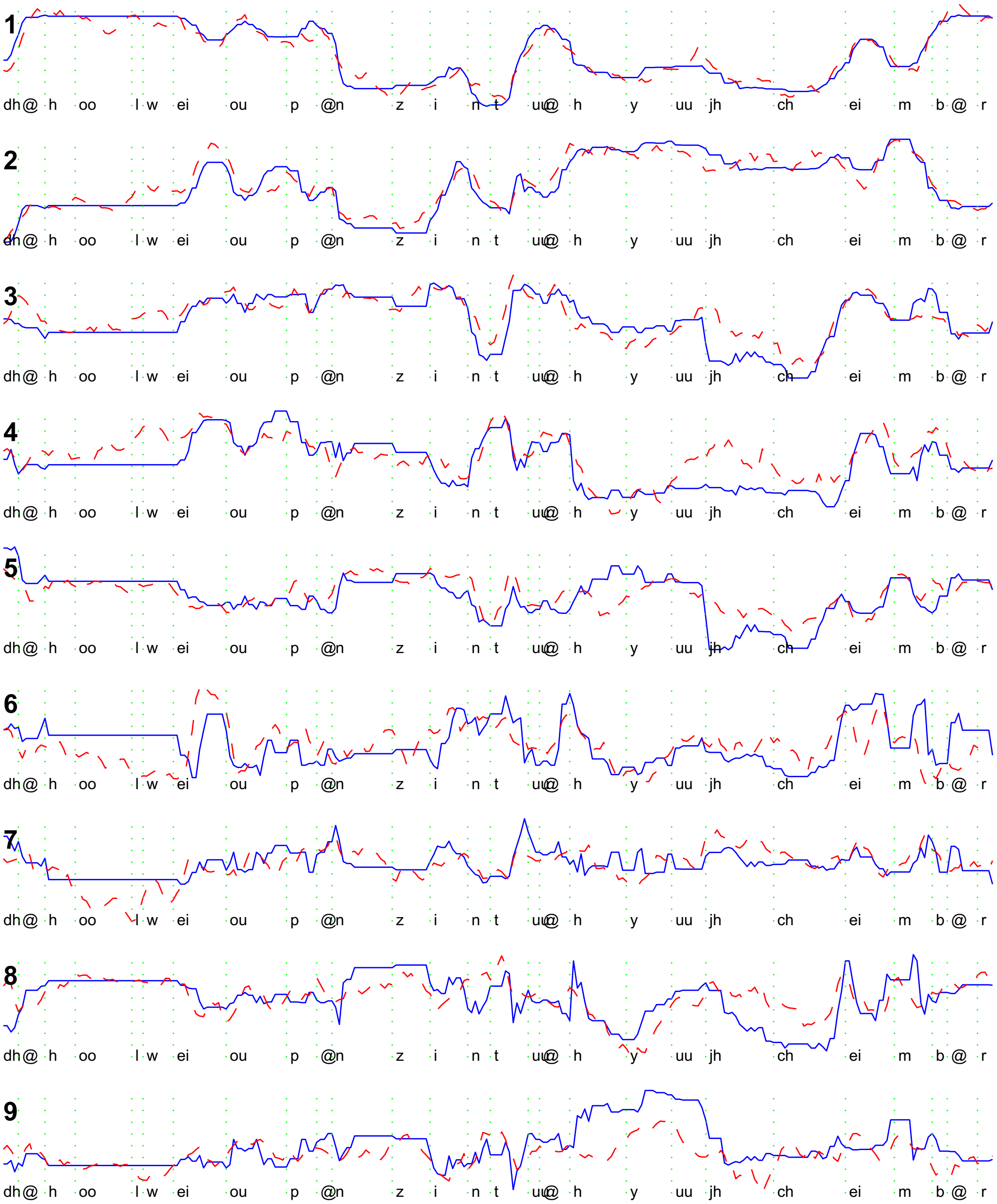

Fig. 9. Actual (solid lines) and estimated (dashed lines) projections on the first nine principal components when fsew0 is uttering the phrase "The hallway opens into a huge chamber". MOCHA labels are shown.

Fig. 11 presents a detailed example of EPG pattern estimation. Again, fsew0 uters the phrase "The hallway opens into a huge chamber". Errors are highlighted by different grey levels according to the legend of the figure. We may identify three kinds of errors. Firstly, there are errors which lead to impossible articulatory configurations. For example, in the l sound, 1st row, 7th pattern, the method outputs a pair of isolated contacted EPG electrodes in the middle of the pattern. The same, with just one contacted electrode, happens in the h sound, 4th row, 3rd pattern. These errors are clearly plain faults of the method used. Secondly, there are some critical errors that alter the overall shape of the vocal tract. For example, in the h sound, 4th row, 4th pattern, and the ch sound, 5th row, 4th and 5th pattern, the method estimates a full palatal vocal tract closure in lieu of

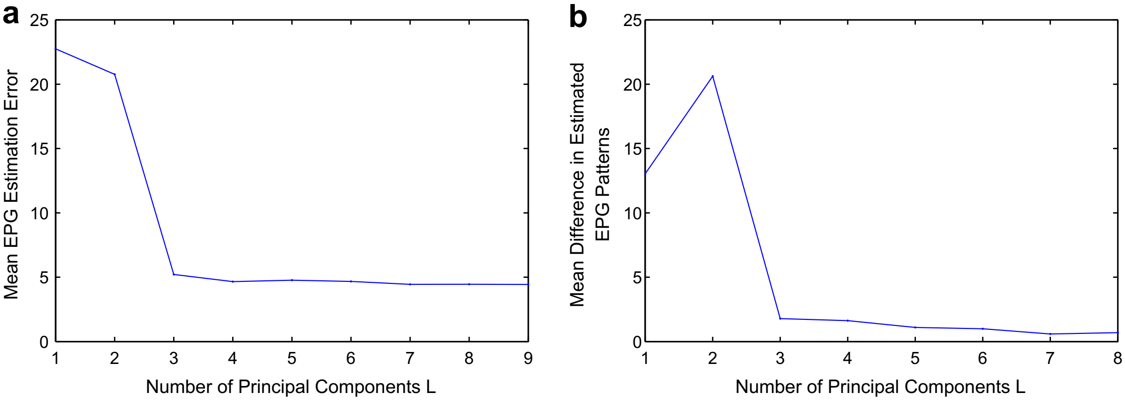

Fig. 10. (a) Reconstruction error as a function of the number of principal components, calculated as mean number of "misclassified" EPG elements; and (b) mean difference between estimations using L and $L-1$ principal components.

Table 2
Numerical summary of the results presented in Fig. 2 (first nine components), Figs. 8 and 10

Principal components	Eigenvalue	Reconstruction error	NMSE	Correlation	Estimation error	Difference
1	8.42	22.34	0.42	0.91	22.75	–
2	4.08	19.72	0.54	0.85	20.77	13.04
3	1.39	4.12	0.71	0.72	5.22	20.62
4	1.17	3.21	0.74	0.68	4.66	1.78
5	1.07	3.36	0.74	0.68	4.77	1.62
6	0.67	3.20	0.88	0.55	4.67	1.10
7	0.64	2.83	0.84	0.59	4.45	0.99
8	0.54	2.80	0.89	0.51	4.45	0.59
9	0.46	2.67	0.84	0.58	4.44	0.69

a relatively open tract. Finally, there are numerous non-critical errors, where the values of some EPG elements are changed without affecting the overall shape of the vocal tract. A characteristic example is the z sound, 3rd row, 3rd pattern, where two EPG elements mutually exchange values, without seriously affecting the overall pattern.

Fig. 12 is analogous to Fig. 6, plotting this time the EPG estimation error against the total number of contacts in each EPG pattern, broken down by phonemic label. Fig. 13 highlights the differences between real and estimated mean EPG patterns across each phonemic label. Examination of these figures shows that EPG pattern estimation is most difficult for phonemes presenting high levels of contact between the tongue and the palate, especially the affricates ch, jh and the postalveolar sibilants sh, zh.

8. Discussion

This paper presented a method for estimated EPG patterns from the speech signal, employing PCA as an intermediate step. Another method to estimate EPG information from the acoustic signal could be to consider the activation of each EPG electrode as a distinct, binary classification problem, independent from the activation of the other electrodes. We pursued this approach in the past with moderate success and compared it with an approach similar to the one presented in this paper, to conclude that the latter performed better (Toutios and Margaritis, 2006).

We chose to work with nine principal components, based primarily on the suggestions of other authors. The results indicate that other choices would also be reasonable. If the mean estimation error of the EPG data was the absolute criterion, we could say, based in Fig. 10a, that four principal components are enough for the task at hand, since the mean estimation error saturates from that point on. On the other hand, Fig. 10b suggests that, at least up to the 9th component, the inclusion of further components in the setup changes,

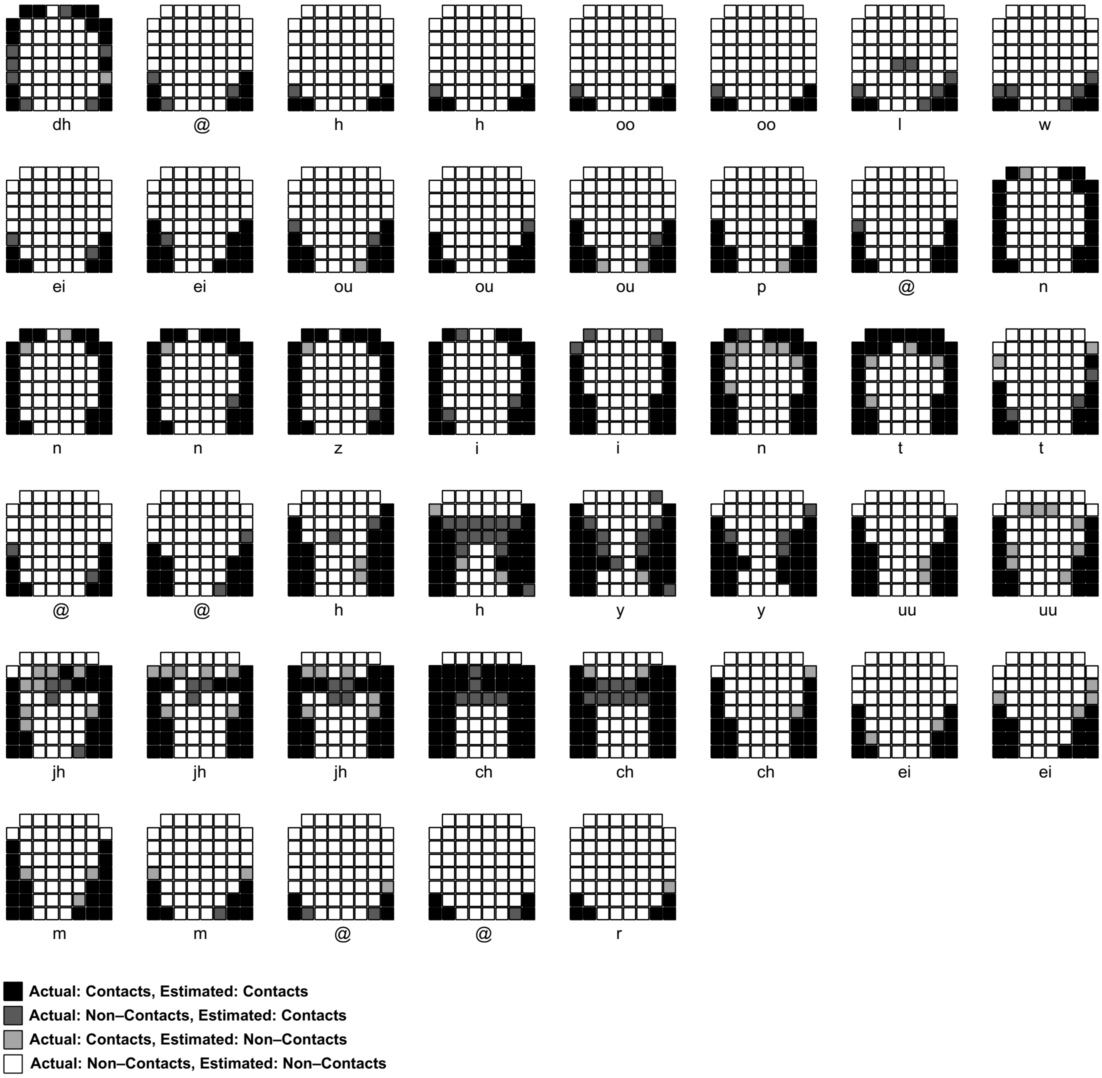
dh
@
h
h
oo
oo
l
w
ei
ei
ou
ou
ou
p
@
n
n
n
z
i
i
n
t
t
@
@
h
h
y
y
uu
uu
jh
jh
jh
ch
ch
ch
ei
ei
m
m
@
@
r
■ Actual: Contacts, Estimated: Contacts
▩ Actual: Non–Contacts, Estimated: Contacts
▨ Actual: Contacts, Estimated: Non–Contacts
□ Actual: Non–Contacts, Estimated: Non–Contacts

Fig. 11. Detailed example of EPG estimation, showing one EPG pattern every 60 ms. Utterance is "The hallway opens into a huge chamber". MOCHA labels are shown for reference.

significantly enough, the produced estimated patterns. Perhaps, the inclusion of further components (probably beyond the 9th) should be stopped when this metric of difference plotted in the specific figure saturates. Anyway, the efficiency of EPG pattern estimation with our method is strictly limited by the efficiency of PCA reconstruction.

We should stress that the experiments presented involved a single speaker. We cannot be absolutely sure, what the implications would be if the experiments were repeated using data from more speakers. The use of more speakers, employing several articulatory strategies, would probably lead to increased mean estimation errors, when estimated EPG patterns are compared to the actual ones.

In several discussions with colleagues, it was suggested to us that the addition of visual cues on lip opening and rounding to the input of our method would probably lead to improved results. We explored this probability in Toutios and Margaritis (2007) employing, however, a different method (a neural network) for estimating EPG patterns. We reported a 6.4% relative decrease in the EPG estimation error when lip opening information is added to the acoustic input and a 7.3% relative decrease with the addition of both lip opening and lip protrusion information.

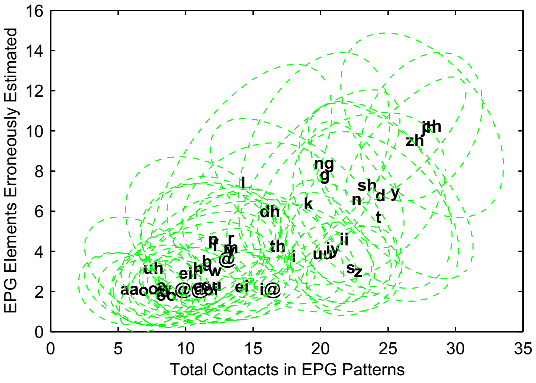

Fig. 12. EPG estimation error plotted against the total number of contacts in EPG patterns and broken down by phonemic label. MOCHA labels are located at the point of the mean value of number of total contacts and estimation error across instances of the labels.

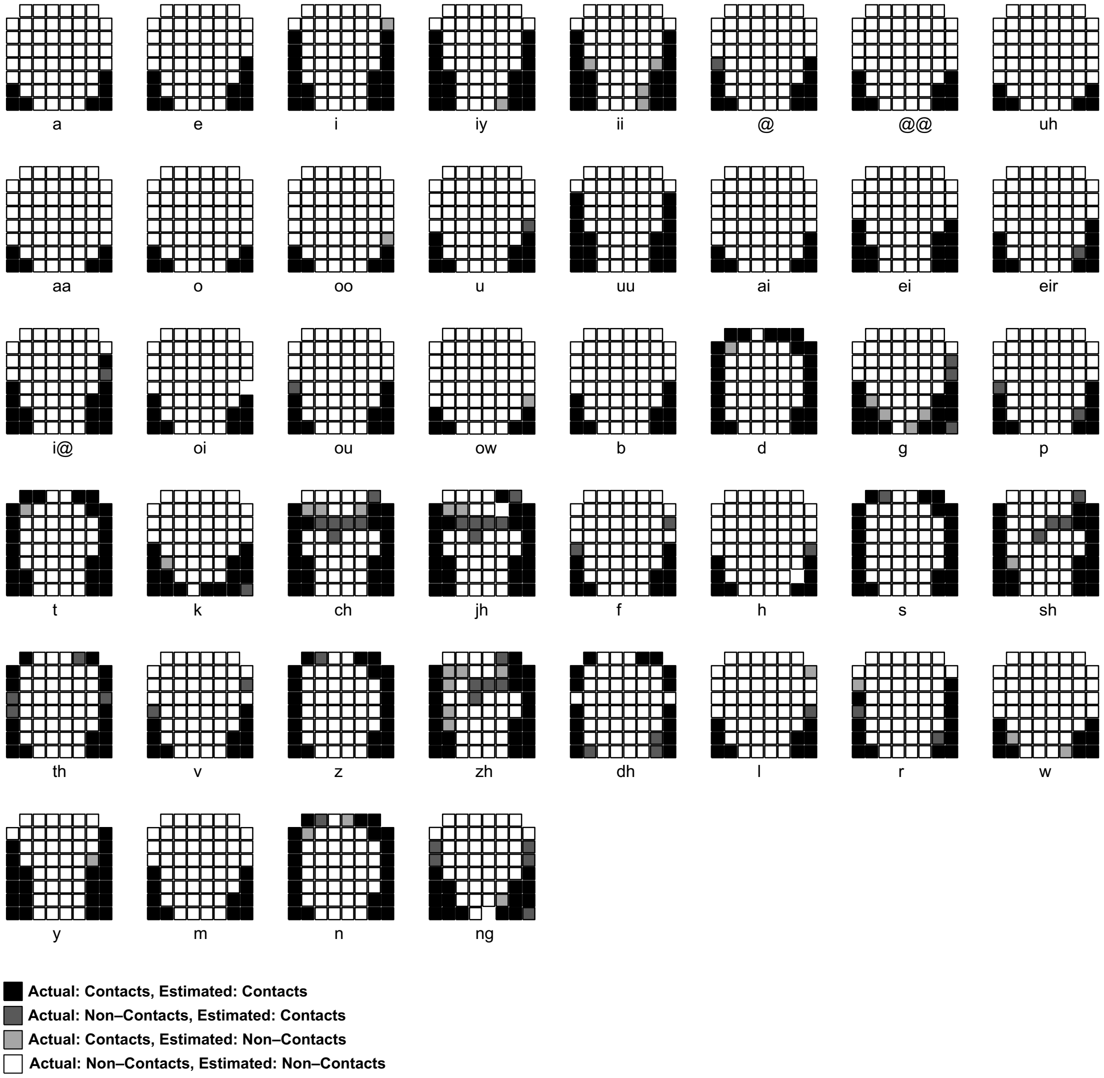
a
e
i
iy
ii
@
@@
uh
aa
o
oo
u
uu
ai
ei
eir
i@
oi
ou
ow
b
d
g
p
t
k
ch
jh
f
h
s
sh
th
v
z
zh
dh
l
r
w
y
m
n
ng
■ Actual: Contacts, Estimated: Contacts
▩ Actual: Non–Contacts, Estimated: Contacts
▨ Actual: Contacts, Estimated: Non–Contacts
□ Actual: Non–Contacts, Estimated: Non–Contacts

Fig. 13. Differences between real and estimated “mean” EPG patterns across MOCHA labels in the test set.

Beyond the estimation of the actual EPG sequences, we feel that an important finding of this paper is the efficiency with which the projections of the EPG data on the first few principal components are estimated from the speech signal. Having in mind several works that use articulatory features (that are possibly estimated through speech inversion) in a speech recognition setup (Deng, 20060, we may assume that the projections on the principal components could be used this way as well. This is suggested by the graphs in Fig. 4 which show clear differences in the values of the projections on these principal components across different phonemes.

References

Carreira-Perpiñán, M.A., Renals, S., 1998. Dimensionality reduction of electropalatographic data using latent variable models. Speech Communication 26, 259–282.

Cherkassky, V., Ma, Y., 2004. Practical selection of SVM parameters and noise estimation for SVM regression. Neural Networks 17, 113–126.

Collobert, R., Bengio, S., 2001. SVMTorch: support vector machines for large-scale regression problems. Journal of Machine Learning Research 1, 143–160.

Deng, L., 2006. Dynamic Speech Models: Theory, Algorithms and Applications. Morgan and Claypool Publishers.

Engwall, O., 2005. Introducing visual cues in acoustic-to-articulatory inversion. In: Interspeech 2005, Lisbon, Portugal, pp. 3205–3208.

Frankel, J., 2003. Linear dynamic models for automatic speech recognition. Ph.D. thesis, The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., 1993. Documentation for the DARPA TIMIT acoustic–phonetic continuous speech corpus CDROM. Tech. Rep., NTIS order number PB91-100354.

Gibbon, F., 2005. Bibliography of electropalatographic studies in English. Tech. Rep., Queen Margaret University College, Edinburgh, UK.

Gibbon, F., Nicolaidis, K., 1999. Palatography. In: Hardcastle, W.J., Hewlett, N. (Eds.), Coarticulation in Speech Production: Theory, Data, and Techniques. Cambridge University Press, Cambridge, England, pp. 229–245.

Hiiemae, K.M., Palmer, J.B., 2003. Tongue movements in feeding and speech. Critical Reviews in Oral Biology and Medicine 14 (6), 413–429.

Jolliffe, I.T., 1986. Principal Component Analysis. Springer.

Jones, W., Hardcastle, W., 1995. New developments in EPG3 software. European Journal of Disorders of Communication 30 (2), 183–192.

Mokhtari, P., Kitamura, T., Takemoto, H., Honda, K., 2007. Principal components of vocal-tract area functions and inversion of vowels by linear regression of Cepstrum coefficients. Journal of Phonetics 35 (1), 20–39.

Nguyen, N., Marchal, A., Content, A., 1996. Modeling tongue-palate contact patterns in the production of speech. Journal of Phonetics 24 (21), 77–97.

Ouni, S., Laprie, Y., 2005. Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. Journal of the Acoustical Society of America 118 (1), 444–460.

Richmond, K., King, S., Taylor, P., 2003. Modelling the uncertainty in recovering articulation from acoustics. Computer Speech and Language 17, 153–172.

Schroeter, J., Sondhi, M.M., 1994. Techniques for estimating vocal tract shapes from the speech signal. IEEE Transactions on Speech and Signal Processing 2 (1), 133–150.

Smola, A., Schölkhopf, B., 2004. A tutorial on support vector regression. Statistics and Computing 14 (3), 199–222.

Toutios, A., 2006. Voice and speech processing and recognition: On the use of stochastic methods for the extraction of phonetic sub-phonemic features from the speech signal. Ph.D. thesis, University of Macedonia, Thessaloniki, Greece, (In Greek).

Toutios, A., Margaritis, K., 2006. On the acoustic-to-electropalatographic mapping. In: Faúndez-Zanuy, M., Janer-García, L., Esposito, A., Satué-Villar, A., Roure, J., Espinosa-Duro, V. (Eds.), NOLISP, Lecture Notes in Computer Science, vol. 3817. Springer.

Toutios, A., Margaritis, K., 2007. Enhancing acoustic-to-EPG mapping with lip position information. In: Interspeech 2007, Antwerp, Belgium, pp. 1374–1377.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, Verlag, New York.

Weston, J., Gretton, A., Elisseeff, A., 2003. SVM practical session (how to get good results without cheating). In: Machine learning summer school, Tübingen.

Woodland, P.C., Gales, M.J., Pye, D., Young, S.J., 1997. Broadcast News Transcription Using HTK. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2. pp. 719–722.

Wrench, A.A., Hardcastle, W.J., 2000. A multichannel articulatory database and its application for automatic speech recognition. In: Proceedings of the 5th Seminar on Speech Production: Models and Data. Kloster Seeon, Bavaria, pp. 305–308.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., et al., 2005. The HTK Book (for HTK Version 3.3), Cambridge University Engineering Department, April.