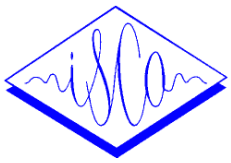

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

ACOUSTIC-PHONETIC FEATURES IN THE FRAMEWORK OF NEURAL-NETWORK MULTI-LINGUAL LABEL ALIGNMENT

Paul Dalsgaard
Speech Technology Centre
Aalborg University, Denmark

William Barry
Dept. of Phonetics
University College London, UK

ABSTRACT

Results are presented from a multi-speaker, multi-lingual method of phonetic label alignment which is based on a combined application of a Self-Organising Neural Network and a Viterbi decoding and level-building technique constrained by an independently specified string of phonetic segments. The Neural Network is trained to convert vectors of cepstral coefficients into vectors of continuously valued acoustic-phonetic features, and to derive a multi-dimensional Gaussian probability density function for each phonemic unit. Multi-lingual application simply requires the definition of the features for each new language. The Viterbi decoding and Level-Building technique is applied to the task of performing label alignment on large speech corpora. The paper firstly presents results for Danish and English, with distributions for selected features and phonemes in the two languages to show the validity of the approach. Covariance analysis within a language allows a reduction of the features to a maximally discriminative set, and comparison across the languages points to the multi-lingual validity of the feature definitions. Secondly, results are given in a number of histograms showing the accuracy of the alignment settings for selected phoneme classes compared to corresponding settings from manually labelled test databases.

The work has been developed in part under the ESPRIT project 'Speech Assessment Methodology' (SAM).

1. INTRODUCTION

A severe problem in the exploitation of phonetic-feature descriptions of a language for Speech Technology purposes is the abstract, categorial manner in which the features are assigned to a given segment. A phonetic segment is only defined with respect to the polarity of the features considered to play a distinctive role in that language; either a feature is present (+), absent (-), or it is not relevant to the sound in question (0). For example, the Danish phoneme /i/ as in "mile" (/mil@/) is in part described by the present vocalic acoustic-phonetic features [+front, +close, +sonorant], and in part by the absent features [-round, -central, -back, -open]. The consonantal acoustic-phonetic features such as [labial, plosive, velar, approximant, nasal etc.] are not relevent to the specification of the sound, and are therefore given as 0. Similarly, the British English phoneme /U/ as in "put" (/pUt/) can be specified by the vocalic features [-front, -central, +back, +round, +close, +mid, -open], and all consonantal features are specified as 0.

Though such categorial feature specifications can operate at the level of abstract structural descriptions, they cannot cater for the variety of articulatory and acoustic manifestations of any given phonetic segment during continuous speech (cf. [1], [2]). Assuming an arbitrary value of +1 for a particular parameter if the feature it represents is optimally present, and -1 if it is maximally absent (i.e. an opposing feature is optimally present), then parameter values can be expected to lie between +1 and -1 (cf. e.g. [3]) as a function of e.g. the speech rate, the degree of stress, and the neighbouring segments.

2. THE LABEL ALIGNMENT SYSTEM

The Label-Alignment system consists of two stages. The first contains a Neural Network, trained to convert a vector of speech signal analysis parameters into a vector of continuously valued acoustic-phonetic features. The second performs Viterbi decoding based on the sequence of feature vectors and one-pass level-building constrained by the given independently specified phoneme string.

The acoustic preprocessing of the speech signal is carried out on the basis of a speech production model, and 12 cepstral coefficients are estimated every 5 msec from a 10 msec window of telephone bandlimited speech. The sampling frequency is 8 kHz.

The cepstral coefficients are fed into a two-dimensional lattice of 20 * 20 hexagonally arranged neurons of a Self-Organising Neural Network. During training this network is firstly exposed to an unsupervised stimulation and a calibration procedure as described in [4], and secondly transformed into a Distinctive Feature Map (DFM). During the DFM-generation each neuron is associated with a vector, each element of which is equivalent to the probability that a specific phoneme is firing that neuron, taking into consideration the a priori probability of the phoneme in the training speech corpus. This allows the transformation of the input cepstral vectors into corresponding acoustic-phonetic feature vectors. This transformation also takes the possibility into account that the neural network was trained with an insufficient number of occurrences of some phonemes. Details are given in [5] and [6].

The output of the DFM is a ϑ-dimensional vector of continuously valued acoustic-phonetic features (ϑ = 20 for the Danish and ϑ = 25 for the British English distinctive feature specified in this paper). A reduction of this dimensionality is carried out on the basis of a full covariance analysis of all features after considering their functionality (results in section 4). On the basis of these 'maximally discriminative features', each phoneme is described by a multi-dimensional Gaussian probability density function (details in section 5).

The stochastically derived phoneme models are used in label alignment, the goal of which is to align a given independently specified phoneme string with the corresponding speech signal taken from a test speech corpus. This is done by means of Viterbi decoding and level-building. The output of the second stage is the optimal alignment positions between adjacent phonetic segments in the pre-specified string.

3. FEATURES AND ACOUSTIC-PHONETIC DATA

The Label Alignment System was applied to two European languages, namely Danish and British English.

The specification of the distinctive features for the Danish vowel and consonant phoneme inventory is given in Table I. The content of this table is utilised during the DFM generation such that phonemes, which are identified within a specific feature-box have their distinctive feature values set equal to +1, in contrast to phonemes of the same broad class (vocalic/consonantal) outside the feature-boxes, which are set equal to -1. Otherwise, their values are set equal to 0.

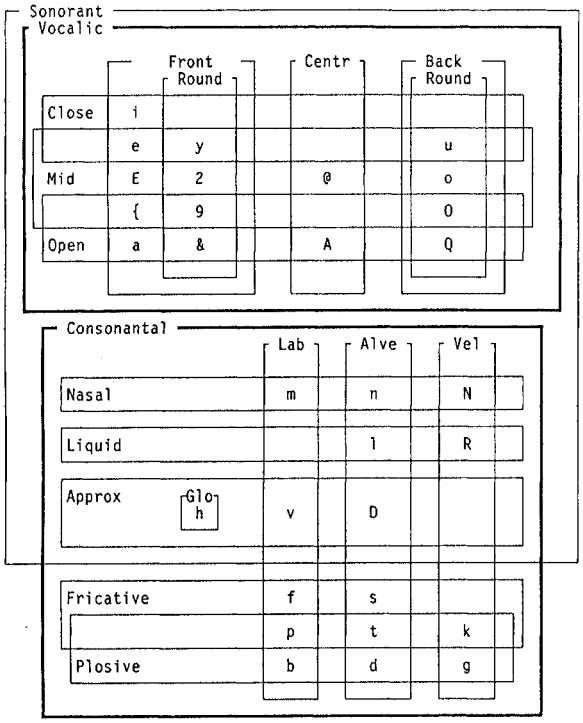

Table I. Danish Distinctive Vocalic and Consonantal Features

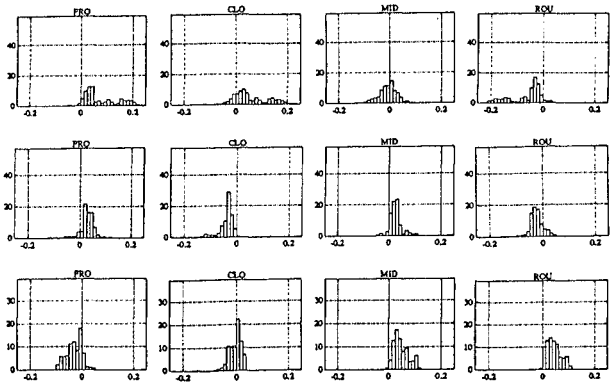

Figure 1. Feature histograms for Danish vowels /i, {, u/

In fig. 1 details of the outcome from the DFM for a limited set of continuously valued acoustic-phonetic features are shown for three selected vowel phonemes /i, {, u/ from top to bottom. Data are taken from the training speech corpus, and given in histograms. The ordinate shows normalised number of occurrences, the abscissa the feature values. Note that, in accordance with the feature definitions for these vowels, [close] and to some extent [mid] differentiate /i/ and /{/; [round], [close] and [front] differentiate /{/ and /u/; [round], [front], and to some extent [mid] differentiate /i/ and /u/.

The specification of the distinctive features for the British English phoneme inventory is given in Table II.

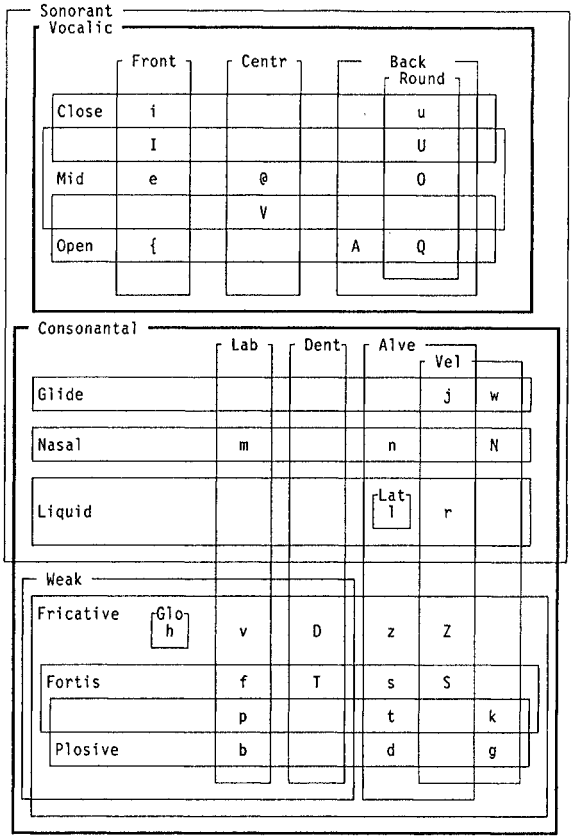

Table II. British English Distinctive Vocalic and Consonantal Features

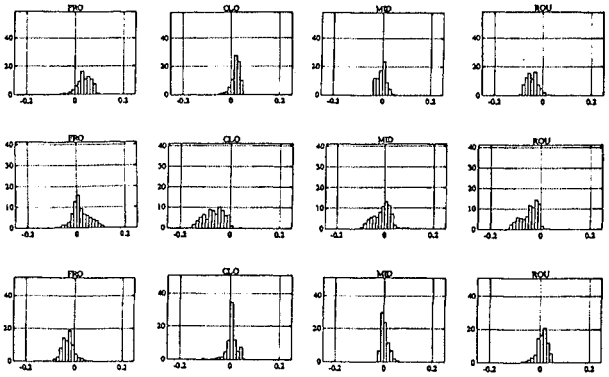

Figure 2. Feature histograms for English vowels /i, {, u/

Fig. 2 shows data from the selected British English phonemes /i/, /{/ and /u/ from top to bottom. It can be seen that there are slight differences in the feature distributions of /{/ and /u/ compared to the corresponding Danish vowels. /u/ does not

manifest the [round] nor the [mid] feature so clearly in British English, and the /{/ is clearly [-close] while being badly defined by [mid] and open (not shown). This lack of clarity in the frame-based feature distribution is the result of individual neurons being shared by data from different phonemic units with opposing feature specification. For example, a neuron reacting to /i/ may also react to /I/ with the feature [+mid], thus giving a distribution containing both [+mid] and [-mid] values for /i/. This may well be due to the slightly diphthongal character of English /i, {, u/, which are often realised with a quality similar to neighbouring vowel (/I, e, U/), specified as [+mid].

4. STOCHASTIC PHONEME MODELLING

For all phonemes a full covariance analysis is carried out on all the acoustic-phonetic features for two purposes. Firstly, to exclude those acoustic-phonetic features, which correlates with a correlation factor $\rho(X, Y) \geq 0.8$ with any other feature from further processing in the Label-Alignment system unless a critical opposition depends on the existence of both features. Secondly, to establish a set of phoneme models, which are all based on a multi-dimensional Gaussian probability density functional description. In Table III (bottom of last page), the full covariance matrix for the continuously valued acoustic-phonetic features is given for Danish. This was used as a basis for feature reduction on condition that the sound (class) represented could be sufficiently discriminated for alignment purposes by other remaining features. The features [consonantal, back, fricative, approximant, glottal] were excluded on the grounds of high covariance with at least one other feature. Following the same criteria, the British English features [vocalic, back, glide1, glide2, dental, lateral] were excluded.

For a specific phoneme p the discriminating function is defined by

$$f_p(\mathbf{x}_d(t)) = L_p^{-1} \cdot \exp(-0.5 \cdot (\mathbf{x}_d(t)-\mu_p)^T \underline{C}_p^{-1} (\mathbf{x}_d(t)-\mu_p))$$

where $L_p = (\| \underline{C}_p \| \cdot (2\pi)^{\vartheta})^{1/2}$ and ϑ is the chosen number of independent features. ϑ = 15 for Danish and ϑ = 19 for British English.

$\mathbf{x}_d(t)$ is the "maximally discriminative" continuous valued phonetic feature vector in frame t. μ_p is the corresponding mean phonetic feature vector and $\underline{C}_p$ the covariance matrix for phoneme p as given from the training process.

5. PHONEME LABEL ALIGNMENT

The set of phoneme models is used during the label alignment. The set is considered as a process, which defines the conditional probability of emitting a phoneme given the feature vector $\mathbf{x}_d(t)$, and given that the phoneme models $f_p(\mathbf{x}_d(t))$ are established from continuous speech similar to the training speech corpus. The accuracy of the label alignment approach is documented with a number of histograms showing the agreement of the alignment settings with the manually segmented and labelled reference test speech corpus, see figures 3-6. Overall alignment accuracy ±20ms is 78% for the British English test material and 57% for the Danish. Figure 3 gives a histogram of the English alignment, with no. of frames deviation on the abcissa and normalised count on the ordinate.

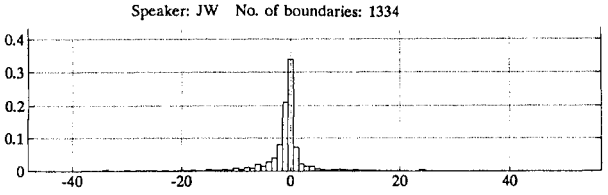

Figure 3. British English Alignment results

There is an overall tendency in both languages for the automatically defined boundaries to be located early in relation to the manually set boundaries. Examining this more closely in the individual class transitions for broad classes of speech sounds in English, it is apparent that the tendency is non-random. Deviations from the manual labels are spread more or less symmetrically around 0 for transitions from sonorants and vowels to obstruents (stops and fricatives), whereas they are negative for transitions from obstruents to sonorants and vowels (see figure 4):

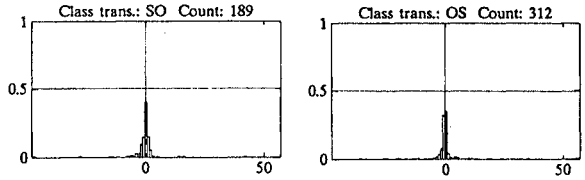

Figure 4. British English alignment between obstruents and sonorants

This asymmetry needs to be investigated further, in particular with regard to the criteria and consistency of automatic and manual boundary placement. Another clear asymmetry is vowel-to-nasal vs. nasal-to-vowel (figure 5), which is explainable in terms of known coarticulation effects in English, anticipatory nasal coarticulation in the prenasal vowel presumably being identified by the system as an early onset of the nasal.

Low accuracy alignment in vowel-to-/l/ and glide-to-vowel transitions (figure 6) is to be expected due to the total lack of boundary in the latter, and the glide-like change in vowels in connection with post-vocalic "dark" /l/, though there is obvious need for more training data as well.

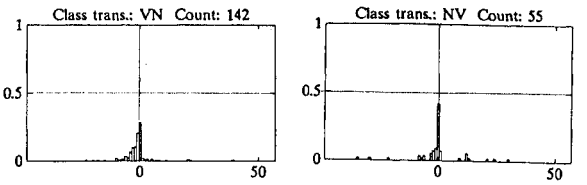

Figure 5. British English alignment between nasals and vowels

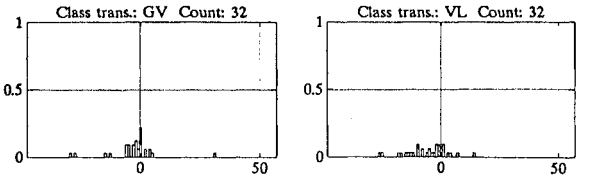

Figure 6. British English alignment for glide-to-vowel and vowel-to-/l/

Similar problems of inherently difficult delimitation are found for the Danish glides and vowel-to-/l/ transition, but also for the Danish approximants, as shown in figure 7. The existence of this class of sounds in Danish in addition to the glides and the vowel-to-/l/ uncertainty explains some of the difference in overall alignment accuracy.

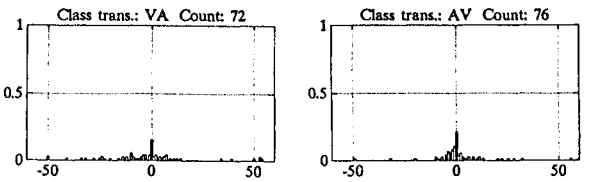

Figure 7. Danish alignment between approximants and vowels

Transitions between vowels and fricatives is extremely accurate, as figure 8 shows, and does not differ appreciably from the accuracy found for English.

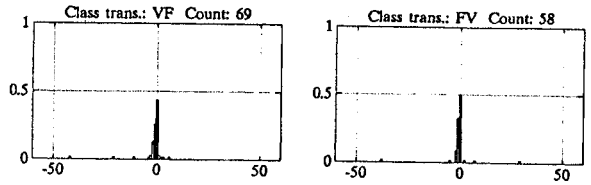

Figure 8. Danish alignment between fricatives and vowels

Transitions between vowels and fricatives are extremely accurately positioned, as figure 8 shows, and does not differ appreciably from the accuracy found for English.

The proportion of boundary placements within ±20ms of the manual label is 57% for the Danish approximants-to-vowels with mean value of 3 ms and standard deviation of 63ms, and 93% for the fricatives-to-vowels with mean value of -15ms and standard deviation of 56ms.

The corresponding ±20ms values for British English are 56% for the glides-to-vowels with mean value of -15ms and standard deviation of 50ms, and 91% for the fricatives-to-vowels with mean value of -8ms and standard deviation of 24ms.

7. CONCLUSIONS AND FUTURE WORK

The result of the alignment comparisons indicate that segmentation was generally encouraging. Some of the boundary discrepancies are, of course, due to the need for better phoneme models based on more data, but it is also clear that improvement for some sounds (e.g. the glides /j/ and /w/, and the approximants in Danish) must remain limited due to the necessary fuzziness of the criteria for delimiting them. The basic theoretical paradox between continuously defined acoustic-phonetic segments and categorically defined abstract phonological units cannot be overcome.

Given sufficient speech data, the refinement of feature definitions should probably be taken in the direction of allophonic differentiation. However, there is also need to address the question of the natural dynamics of speech sounds such as glides and diphthongs, which cannot be resolved on the basis of feature extraction in a single-frame model. Within the wider frame of speech recognition, there is also the problem to be addressed of natural feature redundancy within linguistic systems which leads to high covariance of features of limited functionality outside the particular opposition for which they are required.

To exploit the full power of the feature based approach to multilingual labelling and recognition, the functional status of the acoustically defined features across languages needs to be examined further.

REFERENCES

[1] P. Ladefoged. A Course in Phonetics. London: Harcourt Brace Jovanovich, 1982.

[2] F. J. Nolan. The limits of segmental description. Proceedings of XII International Congress of Phonetic Sciences Vol. 5, 411-414, Tallinn 1987.

[3] R. Jakobson, C. G. M. Fant and M. Halle. Preliminaries To Speech Analysis. The Distinctive Features and Their Correlates. Technical Report, Acoustic Laboratory, June 1955, MIT.

[4] T. Kohonen, K. Torkkola, M. Shozaki, J. Kangas & O. Venta. Phonetic Typewriter for Finnish and Japanese. Proceedings ICASSP88, New York, 607-610.

[5] P. Dalsgaard. Semi-Automatic Phonemic Labelling of Speech Data using a Self-Organising Neural Network. Proceedings EUROSPEECH89, Paris, 541-544.

[6] P. Dalsgaard. Phoneme Label-Alignment using Acoustic-Phonetic Features and Gaussian Probability Density Functions. To appear.

[5] L.R. Rabiner, J.G. Wilpon & F.K. Soong. High Performance Connected digit Recognition, Using Hidden Markov Models. Proceedings ICASSP88, New York, 119-122.

	VOC	SON	CLO	MID	OPE	FRO	CEN	BAC	ROU	CON	LAB	ALV	VEL	GLO	FRI	PLO	APR	LIQ	NAS	SIL
VOC	1.00	0.89	-0.31	0.25	0.02	0.08	-0.55	-0.24	-0.10	-1.00	-0.29	0.58	0.76	0.86	0.40	-0.53	0.85	0.86	0.59	-0.64
SON	0.89	1.00	-0.27	0.27	-0.00	0.03	-0.51	-0.20	-0.05	-0.89	-0.45	0.51	0.68	0.65	0.09	-0.82	0.68	0.68	0.79	-0.76
CLO	-0.31	-0.27	1.00	-0.09	-0.78	0.33	0.22	0.02	-0.12	0.31	-0.01	0.11	-0.33	-0.13	-0.06	-0.00	-0.15	-0.15	-0.03	0.33
MID	0.25	0.27	-0.09	1.00	-0.44	0.11	-0.61	0.08	0.13	-0.25	-0.02	0.11	0.12	0.18	-0.07	-0.17	0.20	0.16	0.21	-0.19
OPE	0.02	-0.00	-0.78	-0.44	1.00	-0.50	0.27	0.14	0.23	-0.02	0.04	-0.18	0.14	-0.07	0.10	0.10	-0.05	-0.02	-0.14	-0.11
FRO	0.08	0.03	0.33	0.11	-0.50	1.00	-0.23	-0.83	-0.85	-0.08	-0.04	0.15	-0.06	0.16	0.05	0.01	0.11	0.06	0.05	0.02
CEN	-0.55	-0.51	0.22	-0.61	0.27	-0.23	1.00	0.18	0.11	0.55	-0.00	-0.08	-0.40	-0.40	-0.00	0.16	-0.39	-0.39	-0.25	0.42
BAC	-0.24	-0.20	0.02	0.08	0.14	-0.83	0.18	1.00	0.90	0.24	0.06	-0.08	-0.12	-0.23	-0.04	0.03	-0.19	-0.14	-0.09	0.18
ROU	-0.10	-0.05	-0.12	0.13	0.23	-0.85	0.11	0.90	1.00	0.10	0.08	-0.05	-0.03	-0.13	-0.04	-0.06	-0.06	-0.01	-0.03	0.06
CON	-1.00	-0.89	0.31	-0.25	-0.02	-0.08	0.55	0.24	0.10	1.00	0.29	-0.58	-0.76	-0.86	-0.40	0.53	-0.85	-0.86	-0.59	0.64
LAB	-0.29	-0.45	-0.01	-0.02	0.04	-0.04	-0.00	0.06	0.08	0.29	1.00	-0.63	-0.31	-0.14	-0.03	0.51	-0.11	-0.20	-0.43	0.43
ALV	0.58	0.51	0.11	0.11	-0.18	0.15	-0.08	-0.08	-0.05	-0.58	-0.63	1.00	0.29	0.63	0.24	-0.43	0.61	0.59	0.55	-0.19
VEL	0.76	0.68	-0.33	0.12	0.14	-0.06	-0.40	-0.12	-0.03	-0.76	-0.31	0.29	1.00	0.77	0.15	-0.32	0.74	0.83	0.52	-0.33
GLO	0.86	0.65	-0.13	0.18	-0.07	0.16	-0.40	-0.23	-0.13	-0.86	-0.14	0.63	0.77	1.00	0.30	-0.27	0.97	0.95	0.55	-0.17
FRI	0.40	0.09	-0.06	-0.07	0.10	0.05	-0.00	-0.04	-0.04	-0.40	-0.03	0.24	0.15	0.30	1.00	0.03	0.23	0.28	-0.29	-0.37
PLO	-0.53	-0.82	-0.00	-0.17	0.10	0.01	0.16	0.03	-0.06	0.53	0.51	-0.43	-0.32	-0.27	0.03	1.00	-0.34	-0.33	-0.79	0.68
APR	0.85	0.68	-0.15	0.20	-0.05	0.11	-0.39	-0.19	-0.06	-0.85	-0.11	0.61	0.74	0.97	0.23	-0.34	1.00	0.94	0.57	-0.19
LIQ	0.86	0.68	-0.15	0.16	-0.02	0.06	-0.39	-0.14	-0.01	-0.86	-0.20	0.59	0.83	0.95	0.28	-0.33	0.94	1.00	0.51	-0.24
NAS	0.59	0.79	-0.03	0.21	-0.14	0.05	-0.25	-0.09	-0.03	-0.59	-0.43	0.55	0.52	0.55	-0.29	-0.79	0.57	0.51	1.00	-0.33
SIL	-0.64	-0.76	0.33	-0.19	-0.11	0.02	0.42	0.18	0.06	0.64	0.43	-0.19	-0.33	-0.17	-0.37	0.68	-0.19	-0.24	-0.33	1.00

Table III. Full covariance matrix for Danish acoustic-phonetic features