# FROM ACOUSTIC SIGNAL TO PHONETIC FEATURES: A DYNAMICALLY CONSTRAINED SELF-ORGANISING NEURAL NETWORK

*Páll Steingrimsson[1], Bent Markussen[1], Ove Andersen[1], Paul Dalsgaard[1], William Barry[2]*
*[1]Center for PersonKommunikation (CPK), Aalborg University, Denmark*
*[2]Institute of Phonetics, University of the Saarland, Germany*

## ABSTRACT

Articulatorily based acoustic-phonetic features are derived from the speech signal via a Self-Organising Neural Network (SONN) using spectral and energy parameters calculated from single windowed segments of the signal, and dynamically constrained by a cost-minimisation procedure enforcing continuity on the basis of features present in the segment. Results of the smoothed feature traces are compared to a previously calculated, unconstrained feature output.

## INTRODUCTION

The identification of the phonetic structure of an utterance in automatic speech recognition is seen increasingly as a hybrid task of combining pattern-recognition expertise with speech science knowledge [1-3]. Just as word recognition had to give way to recognition based on sub-word segmental units (phonemes or allophones) as the demand for ever larger vocabularies increased, so the segmental units have to give way to sub-segmental, parallel properties (features) as the realisation grows that, in normally produced (i.e. continuous) speech, the acoustic properties of a particular sound vary as a function of articulatory dependencies between consecutive segments [2,4,5]. The task of recovering the abstract phonemic structure from the acoustic signal is thus freed from the need to relate the overall acoustic pattern of a stretch of signal to a particular phoneme, and can exploit changes in particular features.

Two aspects of acoustic change are not usually differentiated explicitly:

(i) the fluctuation in spectral structure and consequent feature values during quasi-constant portions of the speech signal (fricatives, stop closures, nasals and laterals, and stressed-vowel centres), and

(ii) information-bearing spectral change (sonorant glides, diphthongs, and place-signalling CV- and VC-transitions.

Attempting to address the second area without dealing with the first is clearly inefficient. In addition, optimized feature extraction based on solutions to the first problem provides a potential basis for correcting for prosodic variation and vowel-to-vowel coarticulation.

This paper presents a method of employing dynamic constraints within a framework of a SONN for smoothing feature traces.

## THE ACOUSTIC-PHONETIC FEATURE ESTIMATOR

The architecture of the feature estimator comprises two modules as shown in Figure 1.

The first, the preprocessing module, calculates, for each 10 ms windowed segment of the speech signal, a set of parameters consisting of 12 Mel-Frequency based cepstral, 12 delta cepstral, 12 delta-delta cepstral coefficients and the corresponding log energy, delta log energy and delta-delta log energy. The calculation takes place every 5 msec. From these parameters, a vector **a** is selected containing the 20 coefficients which maximise the phoneme separability.

The second module converts the vectors from the selected acoustic parameters **a** into acoustic-phonetic features $\phi$ by means of a SONN, the training of which is described in the next section (see [8] for further details).

### The Self-Organising Neural Network

The SONN consists of a number of neurons - 400 are used in this research - which are arranged regularly in a 20 x 20 rectangular structure. Each neuron *n* has assigned to it a vector ***m***, the size and structure of which corresponds to the acoustic parameter vector



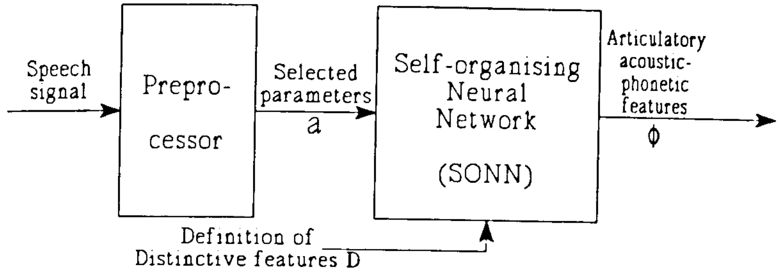


**Figure 1.** Architecture of phonetic feature estimator

**a**, and all neurons are connected in parallel to receive the same input.

The training session comprises three phases. Firstly, an unsupervised stimulation phase in which the SONN input is presented to speech data from the training speech corpus. Secondly, a supervised phoneme calibration phase and thirdly, a supervised acoustic phonetic feature calibration phase.

During the stimulation phase, each neuron n(x,y) of the SONN is assigned a parameter vector **m**(x,y), which is the weighted average of the acoustic vectors **a** firing (see e.g [8]) neuron n(x,y) during the entire stimulation phase.

As a result the SONN organises itself such that: 1) speech sounds that are acoustically close are represented in neighbouring neurons, 2) speech sounds which carry e.g. the same manner feature tend to group together in larger clusters, and 3) different speech sound classes, e.g. vowels vs consonants, are represented by neurons clustering in groups of classes.

### Calibrating the SONN

I) The first calibration phase operates at phoneme level.

During the phoneme calibration phase, the SONN is submitted to the training speech corpus again, and the number of firings $n((x,y)|\phi_j)$ are registered for all phonemes $\phi_j$ and all neurons n(x,y), $x\in\{1 \cdot\cdot N_x\}$ and $y\in\{1 \cdot\cdot N_y\}$ within the SONN.

Given that n(x,y) is the neuron at position (*x,y*), $\mathbf{N}_s(x,y)$, $s\in\{1 \cdot\cdot Q\}$, is a vector each element of which, $p((x,y)|\phi_j) = n((x,y)|\phi_j)/n(\phi_j)$, represents the frequency of occurrence that a specific phoneme is firing neuron n(x,y). The number of times $n((x,y)|\phi_j)$, that neuron n(x,y) is firing given that phoneme $\phi_j$ is present at the input, is calculated during the first calibration phase, and $n(\phi_j)$ is simply the total number of frames in the training corpus representing phoneme $\phi_j$. By employing a clustering technique several vectors may be assigned to each neuron, i.e. Q > 1.

II) The second calibration phase operates at the level of phonological features.

Each phoneme $\phi_j$ is abstractly represented by a phonologically defined distinctive feature vector $\mathbf{D}_{\phi j}$, $j\in\{1 \cdot\cdot M\}$ where M is the number of distinctive features taken into account (observe that vectors **D** are dependent on the language). For example for the Danish phoneme symbol /u/, $\mathbf{D}_{/u/}$ is given by:

$\mathbf{D}_{/u/}$ = [ [+voi] [+voc] [−fro] [−cen]
[+bac] [+rou] [+clo] [−mid] [−ope]
[−con] [0lab] [0den] [0alv] ...
[0fri] [0plo] [0sil] $]^T$

where '+' means feature *present*, '-' means *absent* and '0' means feature *not relevant*.

Based on vectors $\mathbf{N}_s(x,y)$ and $\mathbf{D}_{\phi j}$, $j\in\{1 \cdot\cdot M\}$, a phonetic framework vector $\mathbf{P}_s(x,y)$ is defined for each neuron n(x,y) [8]:

$$\mathbf{P}_s(x,y) \triangleq [\mathbf{D}_{\phi_1}\ \mathbf{D}_{\phi_2} \cdot\cdot \mathbf{D}_{\phi_j} \cdot\cdot \mathbf{D}_{\phi_M}]^T \times \mathbf{N}_s(x,y),$$
$$s\in\{1 \cdot\cdot Q\}$$

where *Q* is the number of phonetic framework vectors assigned to each neuron.

The elements $\mathbf{P}_s(x,y)(k)$ each represent an approximation to the probability that the k'th acoustic-phonetic feature has been involved in the firing of neuron n(x,y).

## SONN FEATURE ESTIMATION

Previously the above expressions were used to estimate acoustic phonetic features directly from the acoustic speech signal on a frame-by-frame basis.

We have recently investigated new principles for estimating these features in which we include dynamic constraints in a Viterbi based minimisation of a chosen cost-function $C(l)$ over a window extending back from the current speech frame $l$. The basic aim is to smooth the fluctuations in the feature values from one frame to the next.

The cost-function $C(l)$ is chosen so as to contain elements which ensure that spectral changes as well as continuity of articulator movement are taken into consideration during the minimisation.

The first element is the summation of the distances $d_s[\bullet]$ between the incoming acoustic vectors $\mathbf{a}$(l-i) and the neuron weight vectors $\mathbf{m}$(x,y,l-i) as calculated over a window of fixed length L frames. This contribution is focused on the spectral differences within the window.

$$C(l) = \sum_{i=0}^{L-1} \left( d_s[\mathbf{a}(l-i),\mathbf{m}(x,y,l-i)] + w \cdot d_f[\mathbf{P}_s(x,y,l-i),\mathbf{P}_s(x,y,l-i-1)] \right)$$

The second summation adds a weighted contribution which is calculated on the basis the distances $d_f[\bullet]$ which represent the differences in the approximated probabilities given by the phonetic framework vectors $\mathbf{P}_s(x,y,l-i)$ and $\mathbf{P}_s(x,y,l-i-1)$ in the window. The factor w is a relative weighting between the two contributions.

Based on the minimisation, the resulting acoustic-phonetic feature vector $\phi$ is defined as follows:

$$\Phi(l-L+1) = \mathbf{P}_s(x,y,l-L+1).$$

## ACOUSTIC-PHONETIC FEATURES

An example of a feature trace as estimated by the above procedure for Q = 2 is shown in Figure 2a on the next page.

The sentence 'pølsevognen stod midt' with the SAMPA transcription /0 p 2 l s @ v Q n s d0 d o D m e d0/ is transformed into phonetic features by applying the delineated approach.

A careful examination of the features illustrated in Figure 2a show a very close correspondence with the traditional definition of the phonemes as given in [8].

The feature traces shown in Figure 2a may be compared to the corresponding traces for the same speech signal as shown in Figure 2b, where the features are derived by the approach which performs the calculations on a frame-by-frame basis.

## CONCLUSIONS AND OUTLOOK

The figures illustrate that articulatorily based features are indeed derivable, and that articulatory and functional features can operate together (see for example *VOC* and *VOI*, *VOI* capturing vocal fold activity, and *VOC* fairly successfully isolating vocalic segments). Also, as examination of Figure 2a indicates, the traces show a) acoustic dependencies between features that are used independently for phonological definition (see for example *BAC* and *ROU*), b) some clear changes in feature strenght during the time course of segments as defined by manual labelling (marked in figures, e.g. *OPE* for /Q/ and *MID, BAC, ROU* for /o/), and c) some carryover of features from the segment where a feature is relevant to where it is not (e.g. some vowel features into /l/ and /n/).

These are, at least in part, indications of articulatory transitions and coarticulation, which are not directly exploitable in a frame-by-frame system. The smoothed traces thus also provide a diagnostic base for the identification of phonetic events and features which require more dynamically oriented acoustic processing.

It is expected that the smoothed traces will provide a sounder basis for the estimation of segment boundaries and the identification of segments. Future work includes testing on two tasks which have been used previously to demonstrate the usability of the approach, namely that of automatic speech signal label alignment and that of phoneme recognition.

## ACKNOWLEDGEMENTS

This work was partly funded via the support to CPK from the Danish Technical Research Council and partly supported by the Human Capital and Mobility Network project SPHERE (contract CHRX-CT93-0098).

## REFERENCES

[1] Moore, R.K. (1993), " Whither a theory of speech pattern processing", Proceedings of European Conference on Speech Communication and Technology, pp 43-47.



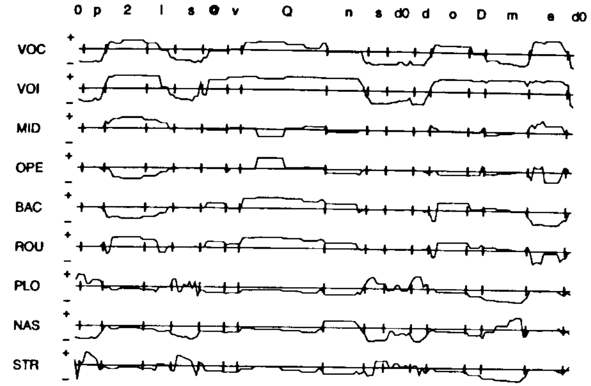


**Figure 2a.** Acoustic-phonetic features derived by the dynamically constrained approach

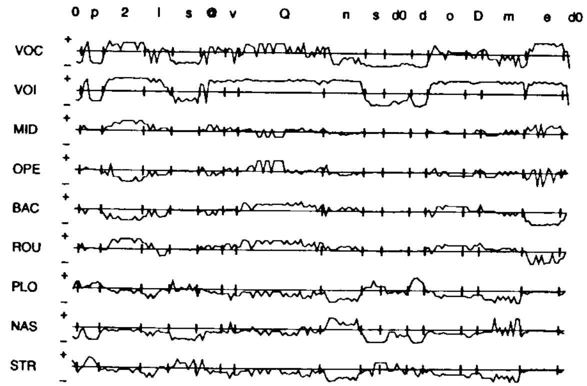


**Figure 2b.** Acoustic-phonetic features calculated on a frame-by-frame basis

[2] De Mori, R. & Flammia, G. (1993), "Speaker-independent consonant classification in continuous speech with distinctive features and neural networks", Journal of Acoustic Society of America 94(6), pp 3091-3103.

[3] Rose, R.C., Schroeter, J. & Sondhi, M.M. (1994), "An investigation of the potential role of speech production models in automatic speech recognition", International Conference on Spoken Language Processing 94, Yokohama, S12-1, pp 575-578.

[4] Deng, L. & Sun, D.X. (1994), "A statistical approach to automatic speech recognition using atomic speech units constructed from overlapping articulatory features", Journal of Acoustic Society of America 95(5), pp 2702-2719.

[5] Deng, L. & Sameti, H. (1994), "Automatic speech recognition using dynamically defined speech units", International Conference on Spoken Language Processing 94, Yokohama, S36-11, pp 2167-2170.

[6] Barry, W. & Dalsgaard, P. (1993), "Speech-database annotation. The importance of a multi-lingual approach", Proceedings of European Conference on Speech Communication and Technology, Berlin, pp 13-20.

[7] Dalsgaard, P., Andersen, O. & Barry, W. (1991), "The cross-language validity of acoustic-phonetic features in label alignment", Proceedings of the XIIth International Congress of Phonetic Sciences, Aix-En-Provence, volume 5, pp 382-385.

[8] Dalsgaard, P. (1992), "Phoneme label alignment using acoustic-phonetic features and gaussian probability density functions", Computer Speech & Language, **6**, pp 303-329.