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

Páll Steingrimsson¹, Bent Markussen¹, Ove Andersen¹, 
Paul Dalsgaard¹, William Barry²

¹Center for PersonKommunikation (CPK), Aalborg University, Denmark 
²Institute 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. The acoustic properties of a particular speech 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 (Fractives, stop closures, nasals and laterals, and stressed-vowel centres), and (ii) information-bearing spectral change (sonant glides, diphthongs, and place-signalling CV and VC-transitions). Attempts 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 φ 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 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]) 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
1) The first calibration phase operates at phoneme level.

The second calibration phase operates at the level of phonological features. Each phoneme φ is abstractly represented by a phonologically defined distinctive feature vector Dφ, je [1 · 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 /l/, Dφ is given by:

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

Based on vectors Nn(x,y) and Dφ, je [1 · M], a phonetic framework vector Pn(x,y) is defined for each neuron n(x,y) [8]:

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

The elements Pn(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 $t$. 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 sum of the distances $d(l-i)$ between the incoming acoustic vectors $a(l-i)$ and the neuron weight vectors $m(x,y,l-i)$ as calculated over a window of fixed length $L$ frames. This contribution is focused on the spectral differences within the window.

$$C(l) = \sum_{i=0}^{L-1} (d(l-i),m(x,y,l-i)) + w \cdot d(P(x,y,l-i),P(x,y,l-i-1))$$

The second summation adds a weighted contribution which is calculated on the basis of the distances $d(l-i)$ which represent the differences in the approximated probabilities given by the phonetic framework vectors $P(x,y,l-1)$ and $P(x,y,l-1)$ 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:

$$\Phi(l+1) = P(x,y,l+1).$$

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 'polsevognen stod midt' with the SAMPA transcription /0 p 2 l s @ v Q n s d D d o D m e d/ 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 OPE for /l/ and MID), BAC, ROU for /l/, 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 /h/).

These are, at least in part, indications of articulatorily based transitions and coarticulation, which are not directly exploitable in a frame-by-frame system. The smoothed traces 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


