THE CROSS-LANGUAGE VALIDITY OF ACOUSTIC-PHONETIC FEATURES IN LABEL ALIGNMENT

Paul Dalsgaard[1], Ove Andersen[1] and William Barry[2]

[1] Speech Technology Centre, Aalborg University, Denmark
[2] Dept. of Phonetics, University College London, United Kingdom

ABSTRACT

Results are reported from a contrastive study in which the validity of articulatory based acoustic-phonetic features were analysed across Danish, English and Italian. The features are derived by means of a Self-Organising Neural Network, trained and calibrated solely on Danish training data. Test material from each of the languages was then used to obtain language-specific distributions for corresponding articulatory features. These were subsequently used to model allophones in the three languages.

Inter-language dependencies in the allophone models were examined in a label-alignment task using techniques from speech recognition to position allophone-transistion boundaries of large speech corpora. The results are stated in terms of accuracy relative to manually placed boundaries.

The work is in part funded by the Danish Technical Research Council and in part by the ESPRIT project 2589, Speech Assessment Methodology (SAM).

1. INTRODUCTION

During the last decade there has been a growth of interest in modelling sub-word units and their phonetic feature definition. This interest stems from the assumption that such an approach to continuous speech recognition will limit the need for large amounts of new speech data each time a new vocabulary is defined. In addition, the feature-based approach takes into account the co-articulation effects inherent in natural speech, and enables the use of well-established search techniques.

'Phonetic-unit' modelling, using triphone models together with whole-word models of function words has been successfully used in e.g. the SPHINX speaker-independent recognition system [1]. Allophonic modelling has been used in a HMM approach [2], where the individual models serve to identify the string of allophones constituting single words. In both approaches the models are based on cepstral coefficients.

In the research presented here, we have developed a new approach in which cepstral coefficients are transformed into a set of articulatory based features by a Self-Organising Neural Network (SONN), and subsequently used to model the 'phonetic units'. The approach has been applied in a continuous-speech recognition system [3], and in the task of label alignment of large speech corpora [4,5]. The latter task was chosen because of the urgent need for labelled speech databases to use in the training of more robust 'phonetic unit' models.

Previously, we have worked individually with a number of European languages. In this paper we examine the cross-language validity of a set of features by applying them to three languages in a cross-language label alignment task.

2. ACOUSTIC-PHONETIC FEATURES

We will only present the main characteristics of the process of transforming cepstral coefficients into a set of articulatory features (e.g. frontness, backness, closeness, dentalness and fricativeness to mention a few). Details can be found in [6].

The technique used is that of a Self-Organising Neural Network [7] consisting of a structure of 20*20 neurons. During the stimulation process each neuron is assigned a vector of adapted cepstral coefficients, and partly due to properties of the SONN and partly to the updating strategy used during the stimulation process, the cepstral vectors describing individual allophones organise themselves into 'clusters' which cover the network in an

orderly way. Acoustically different allophones of the same phoneme will stimulate different neurons in the network.

Following the stimulation process, the entire reference training database is once again presented to the SONN for the purpose of calibration, and the number of firings of each neuron associated with each individual phoneme is counted. The normalised vector of counting rates for each neuron is multiplied with a matrix describing the inventory of phonemes for the language under investigation in terms of a set of distinctive phonetic features [8] giving a vector Φ of acoustically based feature values. The absolute value of each element of Φ is equivalent to the probability that the corresponding acoustic-phonetic feature is involved in the articulation process causing the specific neuron to fire. A neuron fires when the speech frame cepstral vector is the closest to the adapted cepstral vector of the firing neuron taken over all neurons of the SONN.

In each speech frame t, the SONN output is a vector $\Phi(t)$ which in principle could be used as the basis for modelling the individual allophones by a multi-dimensional probability density function [4]. However, some of the features are highly correlated, and therefore the features are submitted to a Principal Component Analysis, the output $\beta(t)$ of which is subsequently used for modelling the allophones. Details may be found in [5].

3. MODELLING OF ALLOPHONES

Each allophone j is modelled by the multiparameter function

$$f_j(\beta(t)) =$$

$$L_j^{-1} \bullet \exp(-0.5 \bullet (\beta(t)-\mu_j)^T \underline{\Sigma}_j^{-1} (\beta(t)-\mu_j))$$

where $L_j = (\|\underline{\Sigma}_j\| \bullet (2\pi)^{\vartheta})^{1/2}$, ϑ the number of parameters in β, $\underline{\Sigma}_j$ the covariance matrix and μ_j the average vector for allophone j as given from the training data.

4. LABEL ALIGNMENT SYSTEM

The functionality of the Label Alignment System (LAS) is based on the assumption that the speech production process can be considered as a stochastic process emitting a sequence of parameters $\beta(t)$, and that the speech signal subsequently being submitted to label alignment manifests the same stochastic behaviour as used during the SONN stimulation and calibration. This allows the LAS to be implemented using the Viterbi Search and Level Building technique, known from speech recognition. In the context of label alignment however, the search is constrained by an independently given string of allophones corresponding to the speech signal being labelled. Details are given in [6].

In previous work we have established three independent LAS's for the European languages Danish, English and Italian, and tested them on large speech corpora. The test have shown the following overall tendencies: I) Labelling accuracy was encouragingly high overall, despite the limited amount of training data. II) Performance was no lower when the LAS were used in 'multi-speaker' mode (i.e trained on different speakers from those used for testing). III) Performance for different sound classes varied according to their acoustic segmentability in a manner that parallels human labelling performance: Fricatives and Plosives were labelled most accurately, post-Vocalic Liquids least accurately.

The results, and the general theory of distinctive features as a language-independent communicative framework, prompted the cross-language experiment reported here. We chose the Danish LAS trained on Danish, and ran the following experiments: A) The LAS was trained and calibrated on the complete reference recording by one male Danish speaker (approx. 2.5 minutes of continuous speech taken from the large, manually labelled reference speech database SAM-EUROM0), and the recordings of one English, one Italian, and one other Danish speaker were used to derive distributions for the common features, and to test the alignment accuracy. B) The LAS was trained and calibrated on 3 Danish speakers (approx. 6.5 minutes of speech) and the recordings of one English, one Italian, and one other Danish speaker were used to derive distributions for the common features, and to test the alignment accuracy.

The rationale for the experiments was 1) to investigate the degree to which cross-language variation in the acoustic expression of the common features affects labelling accuracy, compared to the effect of cross-speaker variation within one language, and 2) to investigate the effect of a larger amount of training data (with accompanying increased variation due to inter-speaker variation) on intra-language and cross-language label alignment.

The implications of the results are both practical and theoretical. In practical terms, cross-language application of a LAS can greatly alleviate the development of recognition systems for other languages by speeding up the annotation of large speech corpora needed for training and research. Theoretically, the results will indicate to what extent distinctive-features can be considered a substance-based system of phonetic distinctiveness that transcends individual language systems, or how far they should be understood as an abstract scheme with no substance outside the individual language.

5. RESULTS

Examination of the feature distributions for the three test speakers (Danish, English, Italian) can be undertaken from two angles. Firstly, the feature distributions for English and Italian derived from the Danish-trained and calibrated SONN can be compared to those found for English and Italian speakers derived from a SONN trained and calibrated on English and Italian, respectively [4,5,6]. Secondly, in preparation for a cross-language alignment test, the distributions of the English and Italian speakers can be compared to the Danish distributions to ascertain the *phonetically* closest sounds.

Lack of space prevents a comprehensive illustration, but in summary, it was found that the English and Italian distributions derived from a SONN trained on the same language are similar to, though somewhat better defined (i.e. with more extreme positive or negative values for the distinctive features) than those derived from a SONN trained on Danish.

As expected, comparison of English and Italian vowels with Danish shows clearly that very acceptable correspondences exist for some vowels, while others deviate in their feature distributions along the dimensions of known phonetic differences between the languages. Transcription conventions (SAMPA symbols used throughout [9]) do not necessarily provide a satisfactory basis for equivalence. Italian (I) and Danish (D) /i/ show good correspondence, while English (E) /i/ is better matched to D /e/ due to the greater degree of closeness of Danish /i/, see Figure 1.

Other good correspondences are : E and I /e/ with D /E/; I /O/ and E /Q/ with D /Q/; E /O/ with D /O/; I /u/ with D /u/; E /V/ with D /A/. In contrast, only weak matches were found for E /u/, which is not well defined by backness, for E /U, I, @/ and I /a/, which all have very weakly defined features.

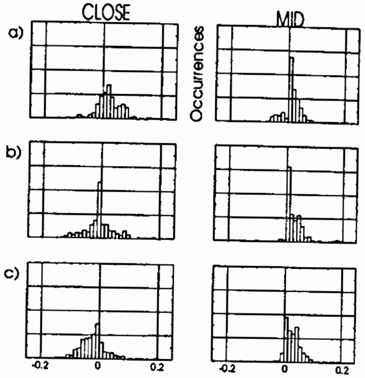

Figure 1. Selected Feature Distributions for a) D /i/, b) E /i/ and c) D /e/

In the consonant systems there are also clear correspondences between Danish and the two other languages (e.g. D, E and I /s/, see Figure 2; D and E /f/; D and E /m/), and cases where only a coarse approximation is possible (e.g. E /S, Z, T/ and I /J, L/).

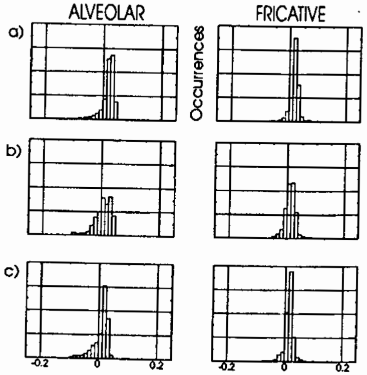

Figure 2. Selected Feature Distributions for a) D /s/, b) I /s/ and c) E /s/

The effect of these approximations is not necessarily manifested in the alignment results. Although overall, alignment accuracy is lower than when English and Italian material is used for training the SONN, greater inaccuracies are not always predictable from the lack of phonetic correspondence. For English, accuracy (to within ±20ms of the manual label) is comparable to the Danish speaker: D 62%, E 60.5%. But this is lower by 15-17% than with a SONN trained on English [4,5]. The most accurate segment transitions are E /s/ and E /S/ together (87%), which are both equated with D /s/, and E /f/ and E /T/ together (79%), which are equated with D /f/. The 47 % overall accuracy rate for Italian at ±20ms is 8% lower than after training with Italian data.

After training and calibration the SONN on three Danish speakers and testing alignment accuracy on the same 3 test speakers as above, accuracy (±20ms) for the Danish speaker rose to 68%, while accuracy for the English and Italian speaker fell to 58% and 46% respectively.

6. CONCLUSIONS

This study illustrates the application of contrastive phonetic principles within a quantitive, speech technology frame. The qualitative assessment of neural-net feature distributions provided a basis for specifying cross-language equivalents in three languages for use in a label-alignment system trained on only one of the languages.

The results of the feature comparison and label-alignment across the three languages indicate that the language-specific manifestation of common features differs enough to make the cross-language application of their distributions less efficient than the language specific application. This is less apparent with the smaller amount of training data (from one speaker), but becomes increasingly evident when the SONN is trained on 3 Danish speakers, given it greater coverage of natural variability, but making it more language-specific.

Given the obvious importance of covering as much of the natural variability as possible, which the effect of the increased training database showed, and given the relative success of some combined categories (E /s/ + E /S/ together, and E /f/ + E /T/ together), a modified approach to multi-lingual labelling appears feasible.

Future research will need to examine ways of covering inter-language variability over approximate phonetic equivalents while broadening the reference categories for alignment purposes. Combining cross-language sound groups into inter-language "broad-categories" [3, 10] for cross-language training is one way in which this might be achieved.

7. REFERENCES

[1] Kai-Fu Lee (1989), "Automatic Speech Recognition, The Development of the SPHINX System", Kluwer Academic Publishers

[2] A. Ljolje, S.E. Levinson (1991), "Development of an Acoustic-Phonetic Hidden Markov Model for Continuous Speech Recognition", IEEE Trans. on Signal Processing, Vol 39, No 1, pp. 29-39.

[3] P. Dalsgaard, A. Baekgaard (1990), "Recognition of Continuous Speech Using Neural Nets and Expert System Processing", International Journal of Speech Communication 9, pp. 509-520.

[4] P. Dalsgaard, W. Barry (1990), "Acoustic-Phonetic Features in the Framework of Neural-Network Multi-Lingual Label Alignment", Proceeding Int. Conf. On Spoken Language Processing ICSLP90, Nov. 18-22, Kobe, Japan.

[5] P. Dalsgaard, O. Andersen, W. Barry (1991), " Multi-Lingual Label Alignment Using Acoustic-Phonetic Features derived by Neural-Network Technique", IEEE Int. Conf. ICASSP91, May 14-17, Toronto, Canada.

[6] P. Dalsgaard (1990), "Phoneme Label Alignment using Acoustic-Phonetic Features", submitted for publication.

[7] T. Kohonen (1990), "The Self-Organizing Map", Proceedings of IEEE, Vol 78, No 9, pp. 1464-1480.

[8] P. Ladefoged (1982), "A Course in Phonetics", Harcourt Brace Jovanovitch, Publishers.

[9] J.C. Wells (1988), "Computer-Coded Phonetic Transcription", J. International Phonetics Association 17, No 2, pp. 94-114.

[10] K. Elenius, G. Takács (1990), "Acoustic-Phonetic Recognition of Continuous Speech by Artificial Neural Networks", STL-QPSR 2-3, Quarterly Report, KTH, Stockholm.