Experiments with a new method for fast unsupervised speaker adaptation in continuos speech recognition

Nikko Ström
Dept of Speech Communication and Music Acoustics,
KTH, Box 700 14, S-100 44 STOCKHOLM
email: nikko@speech.kth.se

ABSTRACT

In this paper, a new method for improving continuos speech recognition (CSR) using speaker adaptation is introduced. The method is designed to give a positive contribution to the recognition even when only a small sample of speech is available. The proposed method is unsupervised in the sense that no information about the speaker is provided except the utterance to be recognised. The recognition is based on an artificial neural network (ANN) working on the phoneme level. The hypothesised speaker characteristics as well as a parametrisation of the acoustic segment are input to the ANN. This unified approach makes it possible to search for the optimal speaker parameters and the optimal phoneme sequence in the same optimisation procedure. The effect of speaker adaptation is achieved by assuming that the speaker parameters are constant over the utterance.

INTRODUCTION

It is well established that speaker adaptation can contribute considerably to the performance of automatic continuos speech recognition systems (CSR) (Blomberg 1990), (Huang & Lee 1991). However, the procedures used, often require long adaptation time. Different kinds of supervision is also often required. In some of the possible CSR-applications, these requirements are unwanted. One example is systems for information retrieval with a large number of users (Blomberg et al 1993). The proposed method is an attempt to avoid these requirements and to give a significant adaptation effect even when the first utterance of a new speaker is recognised.

To perform unsupervised speaker adaptation it is necessary to formalise knowledge of speaker differences. In this paper the characteristics of a speaker are described by a set of speaker parameters. These parameters will always be related to the speaker's formant space (for example mean F1).

The problem could be divided into two sub-problems: (i) Given an utterance, estimate the speaker parameters. (ii) Given the utterance and the estimated parameters, find the phoneme sequence. Unfortunately, parameters like mean F1 cannot be estimated using statistical methods unless we have a large ensemble of utterances. This would make the algorithm too slow for our purposes. On the other hand, if the correct phoneme sequence was known, then it would be possible to estimate the speaker parameters from a small number of phoneme samples by measuring the deviations from some standard phoneme formant values. Of course this procedure includes supervision and will not be used here.

The proposed solution is to perform (i) and (ii) simultaneously. The algorithm tries to find the optimal combination of speaker parameters and phoneme sequence with respect to some error measurement.

THE ALGORITHM

i) Assume that a database organised such that sequences of phoneme segments constituting utterances can be extracted, is available. Each segment is parametrisised to give a representation of its acoustical properties
ii) Assume that it is possible to estimate speaker parameters for each speaker.
iii) Assume that a speaker-sensitive ANN with both the acoustical- and speaker parameters as input have been trained.
iv) Assume that an n-best list of phoneme sequences for the input utterance is given. The algorithm will reorder this n-best list (Hetherington 1993). This list will have been generated by some speaker independent recognition module.
v) The task is to find the optimal speaker parameters (giving the minimum ANN-error) for each sequence and reorder the list according to this error measurement. To perform the optimisation efficiently, some properties of ANN's and the back propagation algorithm are exploited. In the forward phase of the back propagation algorithm, the output activities of the phonemes are computed as a function of the acoustical- and speaker parameters. The error is then the sum of the square of the deviations from the target-values. Typically, the target value for the correct phoneme is 0.9 and for all other phonemes 0.1. In this agorithm, the error will be minimised by choosing optimal speaker parameters. In the backward phase, derivatives of the error with respect to activations and weights are computed. Here, only the derivatives with respect to the speaker parameters are used.

Thus, after the forward and backward phase, the error and the gradient of the objective function are avilable. Given this information it is possible to use a gradient descent method to find the optimal parameters. In this study a Quasi-Newton method was used (Luenberger 1984). The optimisation is constrained by the condition that the speaker parameters are constant during the utterance. This is equivalent to speaker adaptation.

THE EXPERIMENTS

The TIMIT database is both large enough and has sufficiently many speakers to test the suggested algorithm. In a study by Carlson & Glass (1990) five speaker parameters was measured. The same speaker- and acoustical representation are used in this study. The speaker parameters are mean formant frequencies and amplitudes of the first three formants and standard deviation values for the first two formants. The acoustical parameters are the first three formant frequencies and amplitudes in the beginning, middle and end of the vowel, the transition speed of the

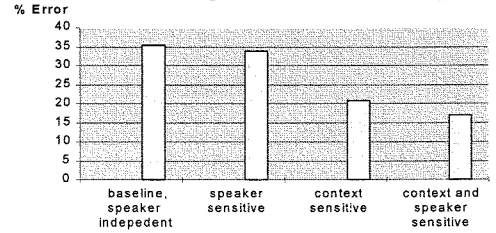

Figure 1. The classification performance in percentage correct classification. The supplied speaker parameters are used where needed.

formants in the three parts and the duration of the vowel (28 parameters). Phoneme sequences are restricted to the sequences of the vowels only. Utterances from 350 speakers were used for training the ANNs and 50 speakers was used for testing.

A speaker independent ANN with only the acoustical parameters as input was trained with the training utterances. Together with a simplified version of the A* n-best search-algorithm (Zue et al 1991) a list of vowel-sequence hypotheses for each test utterance was created.

A speaker sensitive ANN with both the acoustical parameters and the speaker parameters as input was used for the actual speaker adaptation. In some of the experiments a representation of the phonetic context was also input to the ANNs. Classification results for the different ANNs fed with the correct speaker parameters are shown in Figure 1. The results are similar to those reported by Carlson & Glass (1990). Training of both ANNs was performed using the standard back propagation algorithm (Rumelhart et al 1986). All ANNs have 50 hidden units and 16 output units.

The performance of the adaptation depends on the number of vowels in the utterance and the quality of the n-best list. By using the T first vowel tokens from each speaker in the test set, the sequence length can be systematically varied. Performance with varying T is shown in Figure 2. The quality of the n-best list is varied by randomly choosing Q per cent of the segments and impose the correct classification for them. Forcing a

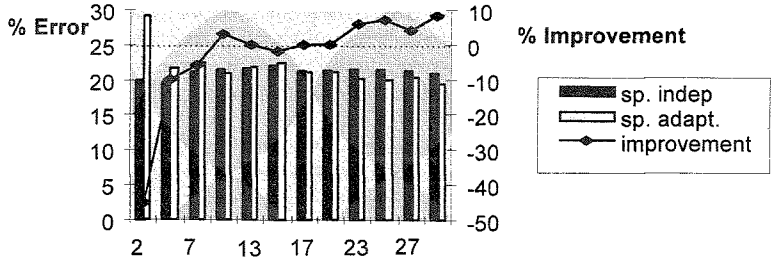

Figure 2. The classification performance for the speaker independent ANN and the results from the speaker adaptation algorithm for different sequence lengths. The curve shows the relative improvement using the proposed adaptation method.

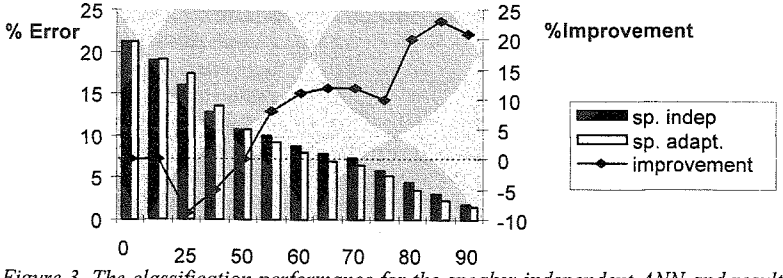

Figure 3. The classification performance for the speaker independent ANN and results from the speaker adaptation algorithm for different values of Q. The curve shows the relative improvement using the proposed adaptation method (T = 20).

percentage of the segments to be correct recognised can be thougth of as a simulation of the effect of incorporating top-down knowledge. Performance with varying Q is shown in Figure 3.

SUMMARY AND CONCLUSIONS

A new algorithm for speaker adaptation has been introduced and tested on the well-documented TIMIT database. The qualitative behaviour of the algorithm is encouraging. As expected, the adaptation contributes more to the recognition performance when the utterance is longer. Also if the quality of the initial phoneme hypotheses is raised, the adaptation contributes more to the performance.

The quantitative results reported from this first study are not fully satisfying. The goal was to reach a significant adaptation effect when only one utterance was presented to the system. From Figure 2. it can be seen that at least 20 vowels is needed to get a positive effect. This is probably longer than most utterances encountered by a CSR-system. It shold be noted that the adaptation performs better with good hypotheses. In a full CSR-system, top-down knowledge contributes to the quality of the hypotheses and therefore the contribution from the adaptation is expected to be increased. In further studies, this expected improvement will be investigated.

ACKNOWLEDGEMENTS

I would like to thank Rolf Carlson and Jim Glass for providing the speaker- and acoustical parameters from their study. The author is sponsored by a donation from VOLVO AB.

REFERENCES

Blomberg, M. (1990): "Adaptatiton To A Speaker's Voice In A Speech Recognition System Based On Synthetic Phoneme References," *Proc. of the tutorial and research workshop on speaker characterisation in speech technology, Edinburgh 26-28 June, 1990*, pp. 58-65

Huang, X.D. & Lee, K.F.. (1991): "On Speaker-Independent, Speaker-Dependent and Speaker-Adaptive Speech Recognition," *Proc IEEE ICASP '91*, pp. 877-880.

Blomberg, M., Carlson, R., Elenius, K., Granström, B., Gustafson, G., Hunnicut, S., Lindell, R., & Neovius, L. (1993): "An Experimental Dialogue System: Waxholm," *Proc EUROSPEECH '93*, pp. 1867-1870.

Hetherington, I.L., Phillips, M.S., Glass, J.R. & Zue, V.W. (1993): "A* Word Network Search For Continuous Speech Recognition," *Proc EUROSPEECH '93*, pp. 1533-1536.

Luenberger, G. L. (1984): *Linear and Nonlinear Programming*, 2nd edition, Addison-Wesley Publishing Company Inc., ISBN 0-201-15794-2.

Carlson, R. & Glass, J. (1990): "Vowel Classification Based On Analysis-By-Synthesis, " *STL-QPSR* No 4, pp. 33-45.

Zue, V et al. (1991): "Integration Of Speech Recognition And Natural Language Processing In The MIT Voyager System," *Proc IEEE ICASP '91*, pp. 713-716.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986): "Learning Internal Representations By Error Propagation," in Rumelhart, D. E., G. E. Hinton, (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1 Foundations.*, chapter 8. Bradford Books/MIT Press, Cambridge, MA, ISBN 0-262-18120-7.