Introducing Phonetically Motivated Information into ASR

Heidi Christensen[†‡], *Børge Lindberg*[†], *Ove Andersen*[†]

[†] Center for PersonKommunikation
Aalborg University, Denmark

[‡] Department of Computer Science
University of Sheffield, UK

hc@cpk.auc.dk, bli@cpk.auc.dk, oa@cpk.auc.dk

Abstract

In this paper we present an approach to introducing more phonetically motivated information into automatic speech recognition in the form of a phonetic 'expert'. To avoid the curse of dimensionality problem, the expert information is introduced at the level of the acoustic model. Two types of experts are used, each providing discriminative information regarding groups of phonetically related phonemes. The phonetic expert is implemented using an MLP. Experiments on a numbers recognition task show that, when using the expert in conjunction with both a fullband and a multi-band system speech recognition performances are increased.

1. Introduction

Within the area of speech recognition the paramount cause of the discrepancies between the performance of humans and machines is the lack of immunity of ASR systems to variations in the acoustic signal not affecting the linguistic message; for example, variations stemming from a change of speaker or speaking style, environmental noise, or channel distortions [1, 2].

Conventional ASR systems rely on a single source of information, which is in sharp contrast to what we know about the way humans process speech. Both physiological and psychoacoustic studies have shown, that human speech recognition is based on several, parallel information extractions from the speech signal [3, 4, 5]. This indicates that incorporating more heterogeneous processing into ASR systems might be a way to leave a possible local performance maximum of current ASR systems.

Relying on multiple sources of information for pattern recognition tasks can increase accuracy and efficiency of the application by taking advantage of inherent weaknesses and strengths of the individual classifiers [6, 7]. The concept of combining classifiers has been analysed lately [8, 9], and within speech applications there has been several studies on the use of sets of classifiers to increase acoustic modelling in speech recognition tasks ranging from large vocabulary speech recognition to classification of syllables, phonemes or groups of phonemes. The *multi-band* framework is one particular type of multiple classifier system that in recent years have proven useful for experimenting with the use of heterogeneous features and information sources [10]. In a multi-band based system adjacent frequency bands are processed independently before being combined at a later stage.

In the area of multiple classifier systems, the crucial question to address is of course: What type of extra information to add? In [11] we showed that combining several standard features in both multi-stream and multi-band type systems could significantly improve performance. In [12, 13] we further showed that designing the feature types so as to be particularly tuned towards a particular phonetic group (such as voiced phones or consonants) also helped improve performance. These experiments have confirmed our hypothesis that introducing in particular phonetically motivated information into ASR can help increase performance, and have encouraged this further work in that direction.

This paper describes another approach also aimed at adding more phonetically motivated information, specifically helping to discriminate between larger groups of phonemes, as confusions between larger groups of phonemes sharing phonetic features are often seen in ASR systems [14]. The information is targeted to resolve some of the confusions and thereby improve performance. What is required is access to a phonetic *expert* that can provide additional, heterogeneous information to an already available, trained system.

The next prominent question is that of how to introduce the expert information into a speech recogniser. One approach when adding information to a statistical pattern recognition system is to simply augment the feature vector with any additional, available features. However, when training statistical models, increasing the dimensionality of the feature space increases the amount of data needed for securing a sufficient estimation of parameters. The phenomenon is often referred to as the *curse of dimensionality*. With a limited amount of training data available other ways of introducing the extra information is of interest. The experiments reported on here avoids the *curse of dimensionality* problem by introducing the additional, heterogeneous information at the level of the acoustic models. The effect of adding the phonetic expert is investigated on two different systems: a conventional fullband system and a multi-band.

The remaining of the paper is organised as follows. Section 2 will briefly present the theoretical formulation of our approach and describe in more detail the training of the phonetic experts. In Section 3 an overview of the data preparation and the used systems are given. Section 4 presents the results from testing the systems on clean and noisy speech, and finally conclusions are given in Section 5.

2. Theory

Assume access to information from an additional classifier, e.g. an expert providing information on the presence of certain phonetic features in the observed data. Conventional ASR systems are defined within a statistical framework and so if the presence of phonetic evidence can be expressed in a statistical manner, i.e. with posteriors, the statistical formulation of the speech recognition problem can easily be modified to accommodate such knowledge.

In ASR the decoding task is aimed at finding the hypothesis or model sequence that is the most likely given the data, $\mathbf{X}$, that

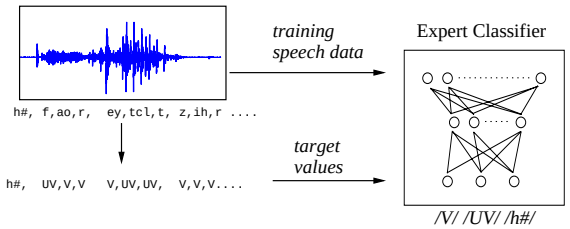

Figure 1: Overview of the training of a phonetic expert MLP. The target values are obtained from the annotations of the database by a direct mapping of the labels, keeping the alignments.

is

$$M^* = \underset{M \in \mathcal{M}}{\operatorname{argmax}} \; P(M|\mathbf{X}) \tag{1}$$

where M is the model sequence, typically word models and $\mathcal{M}$ is the vocabulary used.

The term to maximise when incorporating the expert information into this expression, is then the joint posterior probability of the model sequence M, the system parameters λ and the expert sequence $\mathcal{E}$

$$P(\lambda, M, \mathcal{E}|\mathbf{X}) = P(\lambda, M|\mathbf{X}) \cdot P(\mathcal{E}, M|\mathbf{X}) \tag{2}$$

as λ and $\mathcal{E}$ are independent.

The first term, the posterior probability of the model sequence, is provided by the acoustical model, in this work a multi-layered perceptron (MLP). The second term is in the following work modelled by a separate MLP. This expert MLP is trained to distinguish between a set of expert labels. In the experiments reported here, two different sets of expert labels have been tried out: one classifying speech segments into *voiced, unvoiced* and *silence* and another classifying into five broad phoneme classes, *vowel, consonant, liquid, nasal* and *silence.*

2.1. Implementation of phonetic MLP expert

The phonetic expert MLP is trained similarly to the other MLPs used in the experiments using the same speech data, processed using the same feature processing methods as the fullband system, but with labels mapped to the appropriate target values. Figure 1 illustrates this training process. The expert MLPs have 1500 hidden units, which is the same number as used for the fullband classifier.

3. System overview

The data for training and testing the systems is taken from the Oregon Graduate Institute Numbers95 database of recordings of American English speakers uttering continuous digit and number sequences over the fixed telephone network [15]. 3590 and 1206 utterances from non-overlapping sets of speakers are used for training/cross validation and test purposes respectively. The vocabulary size is 32 words. For testing the noise robustness of the systems, noise samples from the NOISEX database [16] are added per utterance at SNR levels of 0, 6, 12 or 18dB. The *Car noise* and *Factory noise* are chosen for their different spectral characteristics.

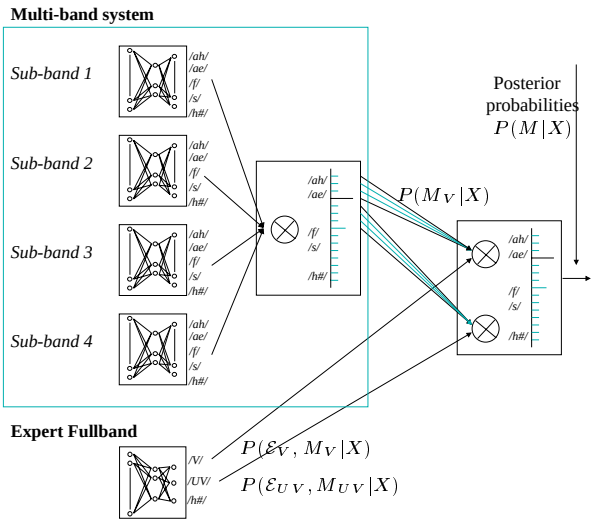

Figure 2: Illustration of the Multiband+VoicingExp system, which is a combination of the multi-stream baseline system with an phonetic expert system trained to perform Voiced/Unvoiced/Silence classifications. The phonetic expert classifiers output the posterior probabilities which are then multiplied to the per-phoneme posteriors obtained from the fullband according to Eq. (2).

Two different feature processing methods are used for extracting basic features plus the energy: Perceptual linear prediction coefficients (plpc) [17] and J-rasta filtered plpc's (j-rasta-plpc) [18]. A feature vector is extracted on 25ms Hamming windowed frames, each overlapping 50%. Delta and delta-delta coefficients (regressing over windows of 5 and 7 frames respectively) are added.

A full-band and a multi-band system, each based on connectionist MLP/HMM entities, are used both individually to provide baseline results and in conjunction with the phonetic expert systems described above. All MLPs are trained on feature vectors derived from 9 frames centered around the current frame and each MLP has 33 outputs representing 32 phonemes and a silence label.

The FullbandBaseline system uses 12 basic features yielding a 39 dimensional feature vector. The MLP has 351 (9×39) input units and 1500 hidden units. The MultibandBaseline system is comprised of four bands with frequency ranges [216-778Hz], [707-1632Hz], [1506-2709Hz] and [2122-3769Hz][1]. 5, 5, 3 and 3 basic features are derived respectively yielding corresponding vector dimensions of 18, 18, 12 and 12. The MLPs have 162 (9×18), 162, 108 (9×12) and 108 input units and 1000, 1000, 660 and 660 hidden units per band respectively. The two baseline systems have a comparable number of parameters.

4. Experimental results

Two series of experiments have been carried out. The first series investigate the effect of adding expert information to a conventional connectionist HMM/MLP system (the Fullband+PhonemeExp and Fullband+BroadExp systems

[1]The frequency bands are chosen so as to roughly capture the formant regions.

respectively). The second series of experiments is concentrated around the baseline multi-band system (Multiband+PhonemeExp and Multiband+BroadExp systems). Each series of experiments is conducted with two types of experts: based on a voiceness criteria or classifying broad phonetic classes. Further each system has been implemented using either plpc or j-rasta-plpc features.

Figure 2 shows one of the four particular configurations, where a multi-band system is combined with a phonetic expert (Voiced/Unvoiced detection) to form the Multiband+VoicingExp system.

4.1. Adding phonetic information to a fullband ASR system

System	j-rasta-plpc	plpc
FullbandBaseline	7.47 %	7.39 %
Fullband+VoicingExpert	6.77 %	7.22 %
Fullband+BroadExpert	7.54 %	6.81 %

Table 1: Word error rates from combining a fullband system with two types of phonetically motivated expert classifiers tested on clean speech.

Table 1 lists the results obtained from testing all fullband based systems on clean speech. It is clear that the systems where phonetic information is added perform better than the corresponding FullbandBaseline systems for the two feature types. This is true in all cases except for the j-rasta-plpc Fullband+BroadExpert system, which performs a little worse than the corresponding FullbandBaseline system. At the same time, the j-rasta-plpc based system, the Fullband+VoicingExpert, is also the feature type that exhibits the best relative improvement.

Figure 3 depicts the word error rates obtained from testing the two systems and the corresponding baseline system in car and factory noise. The figure shows the results from each of the two feature processing methods. For all systems the performance level varies a great deal with the noise type. The car noise is very band limited, and all systems handle it better than the more broad-band and far less stationary factory noise. That being said, the benefits from the extra phonetic information also seem to be noise dependent. When testing with speech to which car noise samples are added the FullbandBaseline performs better for plpc features. For the noise robust j-rasta-plpc the Fullband+VoicingExp systems seems to be slightly better. However, in the case of using factory noise samples the Fullband+BroadExp is consistently better than the other systems for both feature types.

In the above described experiments the behavior of the phonetic expert is crucial for the performance. The experiments here have been designed not so much to achieve an optimal functioning system but rather to demonstrate that adding phonetic information to an ASR system is beneficial. Therefore it was chosen to model the phonetic expert knowledge using an MLP and only limited effort was spent trying to optimise the experts. Nevertheless an improvement in performance was observed when phonetic information was added.

However, one must take into consideration that the phonetic augmented systems described above all comprise approximately twice as many parameters as their corresponding baseline system.

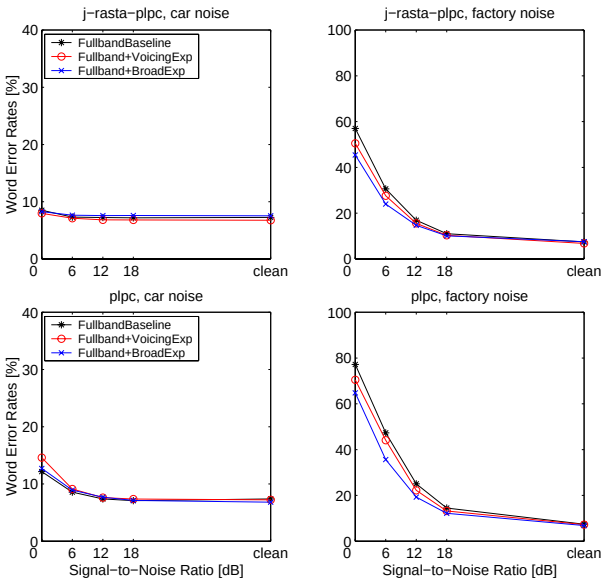

Figure 3: Word error rates plotted against Signal-to-Noise-Ration (SNR) from experiments with systems where a phonetic expert is used in conjunction with a conventional fullband system. The top row and bottom rows correspond to systems based on j-rasta-plpc and plpc features respectively. The left hand side plots are the results from testing in car noise and the right hand side plots presents the results when factory noise is added.

4.2. Adding phonetic information to multi-band systems

Table 2 lists the clean speech results from testing the multiband based systems in conjunction with phonetic experts. The last row in the table is from using a third type of phoneme experts. This last expert is in fact just a normal fullband system. It is clear that vast improvements are obtained and not unexpected the finer and more discriminative the phonetic information is modelled, the lower the WER. In the most extreme case, when for the Multiband+PhonemeExpert (basically a fullband) is added to a conventional multi-band the relative improvements are over 40 % for each of the feature types. This is in accordance with [19].

Comparing the two different feature types, the j-rasta-plpc based systems clearly outperform the plpc based systems. This tendency is likewise apparent when looking at the noisy results, which are plotted in Figure 4, and it can be attributed to the inherent noise robustness of the RASTA filtering [18]. When looking at the results from testing on noisy speech it is

System	j-rasta-plpc	plpc
MultibandBaseline	13.19 %	13.32 %
Multiband+VoicingExpert	11.07 %	11.22 %
Multiband+BroadExpert	10.06 %	10.32 %
Multiband+PhonemeExpert	7.49 %	7.62 %

Table 2: Word error rates from combining a multi-band system with three different experts, the PhonemeExpert corresponds to a normal fullband system, tested on clean speech.

roughly the same pattern as was observed when testing the sys-

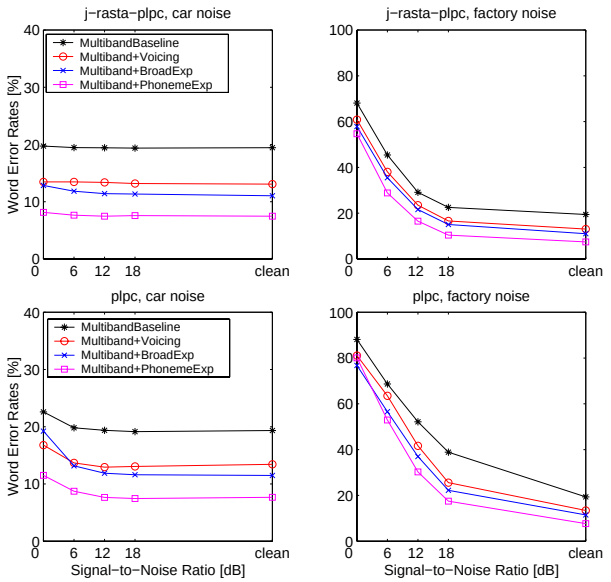

Figure 4: WER plotted against SNR for experiments with systems where a phonetic expert is used in conjunction with a multi-band system. Plots are organised as in Figure 3.

tems on clean speech, namely that PhonemeExpert > BroadExpert > VoicingExpert > multi-band with no expert. For severe SNRs there is a small decrease in how much better the Multiband+PhonemeExpert performs.

Comparing the performance of the multi-band based systems on the two different types of noise it is evident that the system performance degrades far more graceful when operating in car noise.

It is relevant to compare the results for the Multiband+PhonemeExp system to the FullbandBaseline system since the former is the multi-band baseline used together with the fullband and the ladder is the fullband system on its own. For both feature types, the adding of the multi-band system does not help improve clean speech performance despite the fact that the Multiband+PhonemeExp systems have twice as many parameters as the FullbandBaseline systems.

5. Conclusions

In this paper we have presented an approach to introducing more phonetically motivated information into ASR, specifically in combination with a fullband and a multi-band system. We implemented two types of phonetic experts; discriminating between larger groups of phonemes, either Voiced/Unvoiced phonemes or broad phonetic classes. In general we demonstrated an increase in performance when adding the extra information, in particular the multi-band based systems benefited. The experiments indicate that the finer the modelling ability of the phonetic expert, the better the improvement in recognition is.

6. Acknowledgments

This work was done partly as a visiting researcher at the Dalle Molle Institute for Perceptive Artificial Intelligence, Martigny, Switzerland and the Speech and Hearing Group at University of Sheffield, UK. We would like to thank colleagues at both places for valuable discussions.

7. References

[1] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, pp. 1–15, 1997.

[2] L. C. W. Pols, "Flexible human speech recognition," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding 1997*, S. Furui, B.-H. Juang, and W. Chou, Eds., 1997, pp. 273–283.

[3] S. Greenberg, "Auditory function," in *Encyclopedia of Acoustics*, M. J. Crocker, Ed., pp. 1301–1323. John Wiley & Sons, Inc., 1997.

[4] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 1, pp. 115–132, Jan. 1994.

[5] J. B. Allen, "How do humans process and recognize speech," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, Oct. 1994.

[6] T. K. Ho, *A Theory of Multiple Classifier Systems And Its Application to Visual Word Recognition.*, Ph.D. thesis, Department of Computer Science, State University of New York at Buffalo, New York, USA, May 1992.

[7] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[8] K. Turner, *Linear and Order Statistics Combiners for Reliable Pattern Classification*, Ph.D. thesis, Graduate School of The University of Texas at Austin, Austin, USA, may 1996.

[9] L.K. Hansen, "Neural network ensembles," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, Oct. 1990.

[10] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. ICSLP '96*, Philadelphia, PA, Oct. 1996, vol. 1, pp. 426–429.

[11] H. Christensen, B. Lindberg, and O. Andersen, "Employing heterogeneous information in a multi-stream framework," in *Proceedings ICASSP-00*, Istanbul, Turkey, June 2000.

[12] H. Christensen, B. Lindberg, and O. Andersen, "Multi-stream speech recognition using heterogeneous minimum classification error feature space transformations," in *Proceedings NORSIG-00 (IEEE Nordic Signal Processing Symposium)*, Norrköping, Sweden, June 2000.

[13] H. Christensen, B. Lindberg, and O. Andersen, "Noise robustness of heterogeneous features employing minimum classification error feature space transformations," in *Proc. ICSLP '00*, Beijing, China, Oct. 2000.

[14] A. K. Halberstadt, *Heterogeneous Acoustic Measurements and Multiple Classifiers for Speech Recognition*, Ph.D. thesis, Massachusetts Institute of Technology, Massachusetts, USA, nov 1998.

[15] Department of Computer Science and Engineering, "Numbers corpus, release 1.0," Oregon Graduate Institute, 1995.

[16] A. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 CD-ROMs. the NOISEX-92 study on the effect of additive noise on automatic speech recognition," June 1992.

[17] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[18] H. Hermansky, "RASTA processing of speech," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[19] N. Mirghafori and N. Morgan, "Combining connectionist multiband and full-band probability streams for speech recognition of natural numbers," in *Proc. ICSLP '98*, Sydney, Australia, Dec. 1998.