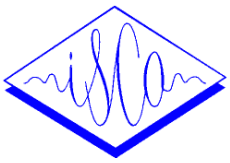


6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

NOISE ROBUSTNESS OF HETEROGENEOUS FEATURES EMPLOYING MINIMUM CLASSIFICATION ERROR FEATURE SPACE TRANSFORMATIONS

Heidi Christensen, Børge Lindberg, Ove Andersen

Center for PersonKommunikation, Aalborg University
Fredrik Bajers Vej 7A, 9220 Aalborg, Denmark
{hc,bli,oa}@cpk.auc.dk

ABSTRACT

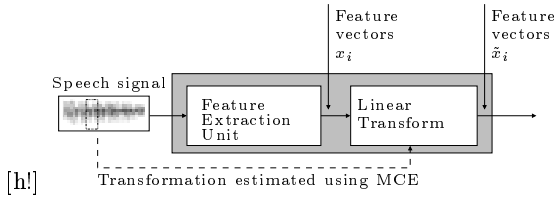
The use of heterogeneous features in automatic speech recognition has been shown to increase clean speech performance. This paper focuses on the noise robustness of systems with heterogeneous features. In particular a system where different features are extracted for different sets of phonemes. The employed features are computed by applying a linear transformation, estimated in a data-driven fashion, to standard feature processing methods. The transformed features are tested in a set of experiments employing different system configurations. Overall the experiments suggests that employing more phoneme specific features can improve speech recognition. When testing the system on noisy speech with added car or factory noise, this tendency is maintained.

1. INTRODUCTION

The performance of an automatic speech recognition system based on statistical techniques is highly dependent upon the quality of the features employed in the system. However, to find a feature extraction method that will allow the extraction of sufficient information to be able to identify the conveyed linguistic message, remains a challenge to the speech community. To obtain close to human performance in a wide range of natural operating environments is a goal far from achieved [13]. In particular when the auditory scene mixes speech with other sound sources, often characterised as 'noise'.

Conventionally just a single feature type is used in a system but recently a series of systems relying on several *heterogeneous* and complementary features have been presented e.g. [2, 7, 8, 11, 15]. Tested on a range of different clean speech recognition tasks each of the systems has been successful. This paper will further investigate how much can be gained from using multiple feature representations by testing the systems noisy conditions. The noisy speech is telephone speech (convolutionary channel noise from various fixed network telephones) mixed at different signal to noise ratios (SNRs) with factory or car noise.

A system relying on an ensemble of classifiers each specialised in classifying within a certain *subset of phonemes* is investigated. Each classifier is based on a unique feature type specifically tuned towards discriminating that particular subset of phonemes, e.g. tuned towards the classification of vowels.



[h!]

Figure 1: Stream specific feature processing.

Several psycho-acoustic and phonetic studies suggest that the use of different features and especially different features for different groups of phonemes can help improve both clean and noisy speech performance. Human speech recognition is based on heterogeneous processing of the signal received by the ear. As a means to convey information speech has evolved in such a way that the different phonetic segments have distinguishable spectral characteristics. The auditory system exploits this by carrying out different processing of e.g. sonorants and non-sonorants [6]. Experiments on the intelligibility of word pairs have shown that different phonetic features are transmitted in different temporal-frequency slots [5]. A related conclusion is made in [16] where it was shown that the optimal frequency range for recognising a phoneme in restricted transmission conditions is very dependent on the phoneme.

1.1. Hierarchical multi-stream framework

One commonly employed architecture when using heterogeneous features is based on the *multi-stream* theory as formulated e.g. in [1]. Fundamental for the paradigm is the use of multiple feature and classifier streams rather than relying on just a single line of feature extraction followed by classification as in more conventional ASR approaches.

Each stream is comprised of a feature extraction unit followed by a classifier, such as a multi-layered perceptron (MLP) network, whose outputs can be considered as posterior probabilities of the observed encoded data. The

probabilities are merged before being used in a conventional hidden Markov model (HMM) state path decoding that produces the recognition hypothesis.

In [2] we demonstrated that with multi-stream systems it is possible to increase clean speech performance by augmenting the heterogeneity of the signal processing employed, even without increasing the number of system parameters. Three rather standard feature extraction techniques were employed but the results encouraged us to exploit the potential advantages of more specifically designed stream processing.

A method for estimating such heterogeneous, tuned features through the employment of a stream specific linear feature space transformation, estimated from speech data using the Minimum Classification Error (MCE) criterion, is presented. The object of the transformation is to enhance the features possessing discriminative information and reduce the effect of the least discriminative ones. An overall research objective in choosing the method has been to find a data-driven solution. The MCE theory and the employed algorithm is briefly described in the following section. The investigated system architecture is described in section 3 and section 4 describes data preparation. Results from applying the estimated transformations on various configurations of the system are presented in section 5 and finally conclusions and directions for future work are given in section 6.

2. REVIEW OF MCE ESTIMATED TRANSFORMATION

The feature processing in each stream is a combination of a standard high-performance feature extraction method and a linear transformation in a matrix form. The transformation matrix is estimated using the Minimum Classification Error (MCE) criterion [4, 12] in an iterative procedure. See figure 1.

In general, a linear transformation of a feature space can be obtained through multiplication with a transformation matrix, U:

$$\tilde{\mathbf{x}} = U(Hx)\,, \tag{1}$$

where H is a normalisation matrix applied to all the data in order to achieve a unit variance distribution for each feature dimension. The elements of the transformation matrix are at each iteration of the estimation procedure computed by *gradient descent* of the *cost function*, L:

$$u_{n,k} = u_{n,k-1} - \eta \left.\frac{\partial L}{\partial u_n}\right|_{u_{n,k-1}}\,, \tag{2}$$

where η is the convergence coefficient. Given a set of training sequences $X_1, \ldots, X_M$ each belonging to a class λ_i from the set of classes $\lambda_i, \ldots, \lambda_I$, the overall cost function is defined as:

$$L = \sum_{m=1}^{M} l_m(X_m) = \sum_{m=1}^{M} \frac{1}{1+\mathrm{e}^{-\alpha d_m(X_m)}}\,, \tag{3}$$

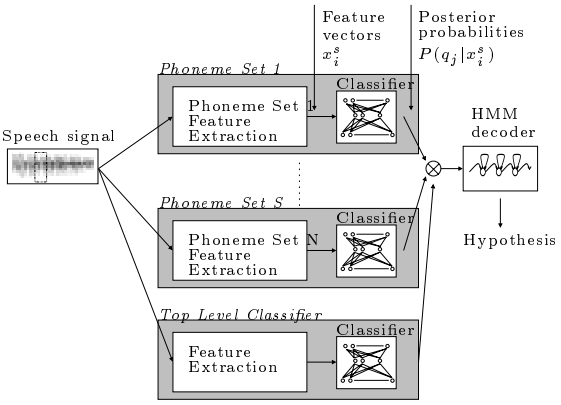


Figure 2: Schematic overview of hierarchical multi-stream based system. Each stream is specialised on a certain group of phonemes.

where $l_m(X_m)$ is the cost function for each training sequence, X_m. Here is used the sigmoid of an error measure, $d_m(X_m)$ defined as:

$$d_m(X_m) = -g_{k(m)} + \frac{1}{\beta}\ln\left[\frac{1}{I-1}\sum_{j\neq k(m)} \mathrm{e}^{\beta g_j}\right]\,, \tag{4}$$

where $g_i = g_i(X_m, \lambda)$ are the discriminant functions and $\lambda_{k(m)}$ is the correct class for the sequence of feature vectors. The discriminant function is defined as the logarithm of the likelihood of the data in sequence X_m

$$g_i(X_m, \lambda_i) \quad = \quad \ln P(X_m|\lambda_i)\,. \tag{5}$$

As in [4] the transformation matrix is trained, on the same training material as is used for training the remaining elements in the system, using a set of simple, spherical Gaussian classifiers to reduce computational complexity. The transformation is then applied in a system with a more complex classifier. However, instead of a Hidden Markov Model (HMM) based system as was used in [4] a hybrid MLP/HMM system is used.

3. ARCHITECTURE

In multi-stream systems such as those investigated in [2], though each stream uses a specific feature processing, it is trained on the *entire training material* and provides posterior probabilities for *all phonemes*. However, in the work presented here the interest is in using heterogeneous features specific to a given subset of phonemes and hence one stream is assigned to each subset. Between them all the streams will of course comprise classification information covering the entire phoneme set, but each stream will only provide posterior probabilities accounting for the phonemes in the given subset. The choice of system architecture is therefore less obvious and in [7] we investigated the clean speech performance of a number of different architectures

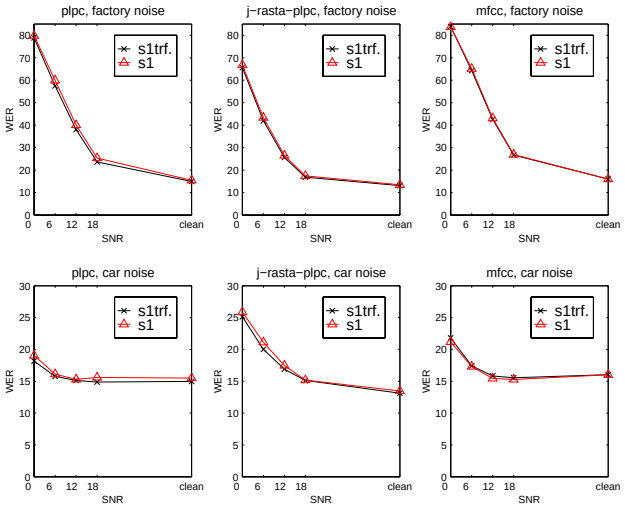


Figure 3: Results of testing with noisy speech using **broad** division when using transformed and untransformed features.

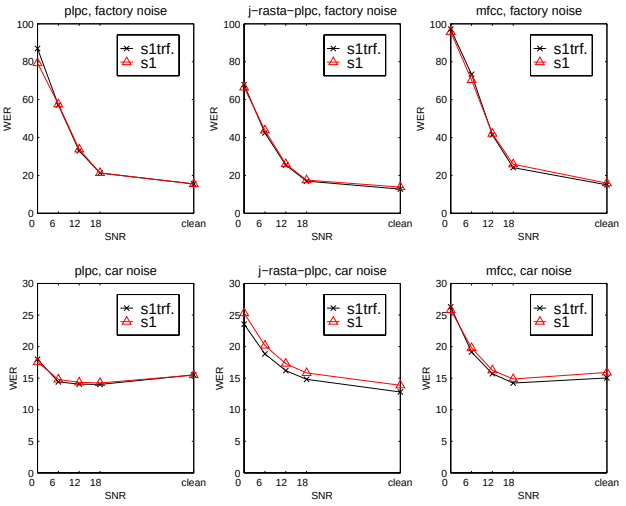


Figure 4: Results of testing with noisy speech using **voicing** division when using transformed and untransformed features.

that support the use of streams with phoneme subset specific processing. The best performing architecture was a hierarchical multi-stream system which is the basis of the experiments reported on here. A schematic overview of the system is presented in figure 2.

The employed system is a hierarchical system with two layers. The **bottom-layer** is a set of classifier experts each specialised in classifying a subset of phonemes. The **top-layer** is an MLP trained at discriminating between the phoneme subsets present in the system. Two different ways of dividing the phoneme set into smaller subsets were tested: either dividing into broad phonetic classes (*vowel, consonant, liquid, nasal, silence*) or, using a voicing criterion, dividing into the classes (*voiced, unvoiced, silence*). With the hierarchical system, using the voicing criterion for dividing the phoneme set, there are three classifier experts handling the *voiced*, *unvoiced* and *silence* phoneme subsets respectively, and the corresponding top-layer MLP is trained to discriminate between the same three subsets. To increase the contextual information of the system, delta and delta-delta features are added to the top-layer MLP. The posteriors from the two layers are combined by multiplying the bottom-layer output with the corresponding top-layer probability.

4. EXPERIMENTAL SETUP

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 [14]. 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 [17] are added per utterance at SNR levels of 0, 6, 12 or 18dB. *Car* and *factory* noise are chosen for their different spectral characteristics.

All streams in each system have their own specific feature processing and classifier units (MLPs). One of three different feature processing methods for extracting the basic features plus the energy: Mel-scale Frequency Cepstrum Coefficient (mfcc's) [3], Perceptual linear prediction coefficients (plpc's) [9] and J-rasta filtered PLPCs (j-rasta-plpc's) [10] are employed. Each feature is extracted on Hamming windowed frames spanning 25 ms, each overlapping 50%. The classifier unit in each stream is an MLP entity trained on feature vectors derived from 9 frames centred around the current frame. Each feature vector is comprised of 12 basic features. The top-level MLPs are trained on basic features augmented with delta and delta-delta coefficients (regressing over windows of 5 and 7 frames respectively) yielding a 39 dimensional feature vector. The number of inputs and outputs of the MLPs varies and depends on the specific stream configuration. All MLPs use 1500 hidden units.

5. EXPERIMENTAL RESULTS

In the series of experiments conducted in this study three different basic feature type (mfcc, plpc or j-rasta-plpc) and two phoneme division (broad or voicing criterion) are tested. To assess the effect of transforming the feature space all experiments are replicated using untransformed features. Results are shown in Figures 3 and 4.

Looking first at clean speech Word Error Rate (WER) in the majority of cases the systems employing the transformed features show a decrease in WER. When analysing the obtained frame scores on the training set, this tendency is even more pronounced. For all pairs of classifier experts from transformed and untransformed features respectively, the experts based on the transformed features all exhibit

better performance than experts trained on the standard features alone [7].

Comparing the different **feature types** (columns of graphs) shows a small but consistent improvement in performance when the features are transformed. The overall performance for all systems degrade when more noise is added. When factory noise is added, as expected the j-rasta-plpc's are doing best, however when using the rather band-limited car noise samples (lower row of graphs), the plpc based systems surprisingly are best. Testing the **noise robustness** of the system show that an increase in performance is maintained when adding either car or factory noise. However, based on these experiments we are unable to conclude that using the suggested heterogeneous features with MCE-based linear transformations increases the noise robustness.

As stated previously an overall objective for these studies has been a search for a data-driven solution, and obviously the pre-defined **phoneme divisions** is a violation against this. A situation we intend to rectify in future work. The voicing division is more coarse and gives slightly more evenly distributed subsets than does the broad class division[1]. It is interesting to see that comparing the results Figure 3 to 4 shows very little difference between the division criteria both when looking at absolute performance level and when looking at relative improvements when using the transformations. Further studies must be conducted to establish to which degree these findings can be generalised.

6. CONCLUSIONS

The underlying hypothesis for the experiments presented in the paper is that by increasing the heterogeneity of the features employed in an ASR system it is possible to increase performance. We have presented a means to obtaining the specific feature extraction methods through applying a linear transformation estimated using the MCE method. A hierarchical multi-stream based system architecture supporting the use of heterogeneous feature extraction for different phoneme subsets, have been tested in different configurations using the transformed features. For a majority of cases the transformation reduces WERs on a numbers recognition task. For all the MLPs trained on the transformed features an increase in per-frame performance was demonstrated. A performance increase is maintained when noise is added to the speech for car and factory noise. Future work will be directed towards system architecture, specifically investigating data-driven ways of choosing the phoneme subsets.

7. ACKNOWLEDGMENTS

This material is based upon research carried out as a visitor of the Speech and Hearing Group at the Department of Computer Science, University of Sheffield.

[1] voiced: 24 phonemes, unvoiced: 9, silence:1 and vowel: 12, consonant: 15, liquid: 4, nasal: 1, silence: 1.

8. REFERENCES

[1] H. Bourlard, S. Dupont, and C. Ris. Multi-stream speech recognition. Technical Report IDIAP-RR 96-07, Dalle Molle Institute for Perceptive Artificial Intelligence, Martigny, Switzerland, December 1996.

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

[3] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-28(4):357, 1980.

[4] A. de la Torre, A. M. Peinado, A. J. Rubio, V. E. Sánchez, and J. E. Diaz. An application of minimum classification error to feature space transformations for speech recognition. *Speech Communication*, 20:273–290, December 1996.

[5] O. Ghitza. Auditory models and human performance in tasks related to speech coding and speech recognition. *IEEE Trans. Speech and Audio Processing*, 2(1):115–132, January 1994.

[6] S. Greenberg. Auditory function. In M. J. Crocker, editor, *Encyclopedia of Acoustics*, pages 1301–1323. John Wiley & Sons, Inc., 1997.

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

[8] A. K. Halberstadt and J. R. Glass. Heterogeneous measurements and multiple classifiers for speech recognition. In *Proc. ICSLP '98*, Sydney, Australia, November 1998.

[9] H. Hermansky. Perceptual linear predivtive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.

[10] H. Hermansky. RASTA processing of speech. *IEEE Trans. Speech and Audio Processing*, 2(4):578–589, October 1994.

[11] A. Janin, D. Ellis, and N. Morgan. Multi-stream speech recognition: Ready for prime time? In *Proc. Eurospeech '99*, pages 591–594, Budapest, Hungary, September 1999.

[12] B-H. Juang and S. Katagiri. Discriminative learning for minimum error classification. *IEEE Trans. Signal Processing*, 40(12):3043–3054, 1992.

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

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

[15] S. Sharma, D. Ellis, S. Kajarekar, P. Jain, and H. Hermansky. Feature extraction using non-linear transformation for robust speech recognition on the AURORA database. In *Proceedings ICASSP-00*, Istanbul, Turkey, June 2000.

[16] H. J. M. Steeneken. *On Measuring and Predicting Speech Intelligibility*. PhD thesis, Instituut voor Zintuigfysiologie-TNO te Soesterberg, Soesterberg, June 1992.

[17] 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.