EMPLOYING HETEROGENEOUS INFORMATION IN A MULTI-STREAM FRAMEWORK

Heidi Christensen, Børge Lindberg and Ove Andersen

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

ABSTRACT

A multi-stream speech recogniser is based on the combination of multiple feature streams each containing complementary information. In the past, multi-stream research has typically focused on systems that use a single feature extraction method. This heritage from conventional speech recognisers is an unnecessary restriction and both psycho-acoustic and phonetic knowledge strongly motivate the use of **heterogeneous** features.

In this paper we investigate how heterogeneous processing can be used in two different multi-stream configurations: first, a system where each stream handles a different frequency region of the speech (a *multi-band* recogniser) and, second a multi-stream recogniser where each stream handles the full frequency region. For each type of system we compare the performance using both homogeneous and heterogeneous processing. We demonstrate that the use of heterogeneous information significantly improves the clean speech recognition performance motivating us to continue exploring more specifically designed stream processing.

1. INTRODUCTION

Using multiple feature streams in automatic speech recognition (ASR) as formulated by Bourlard *et al.* in [2] differs from more conventional ASR approaches in that instead of basing the recognition on a single line of feature extraction followed by classification, multi-stream ASR systems rely on multiple representations of the characteristic information in the speech signal.

The underlying principle of the paradigm is that extracting and fusing diverse information potentially increases the performance, since no error-free solution to the problem exists, and the streams therefore will complement each other in correctness. The overall performance of the system will depend not only on the individual performance level of each stream but also on how well the error patterns of the different streams complement each other. One of the potentials of the multi-stream technique is therefore to fully exploit the redundancy and diversity possessed by the stream specific processing [1].

Figure 1 shows a schematic overview of a general multi-stream system. 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.

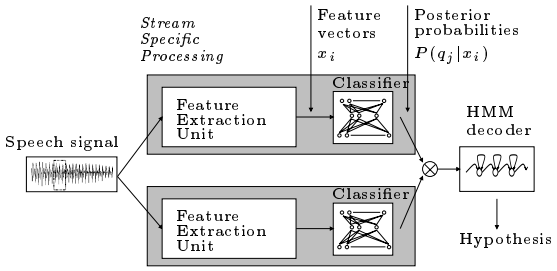

Figure 1: General multi-stream system with two streams.

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

In a multi-stream recogniser the complementariness is achieved through diversity in the streams. In the group of multi-stream systems known as *multi-band* systems, the diversity is enforced by letting each stream handle different frequency regions of the speech signal [2] [3] [14]. Similarly, in *multi-scale* systems each stream operates on different time windows [6] [19].

Another way to augment the complementariness is through the explicit use of heterogeneous features in each stream. Such a principle is closely related to theories governing multiple classifiers and ensemble techniques [13][15], as well as to systems with hierarchical architectures, i.e. [9]. However, within the multi-stream framework very few studies have been presented that focus on increasing the performance through the use of heterogeneous feature processing. Some multi-stream systems have been presented that do make use of different feature processing in each stream, but the effect of the heterogeneity has not been investigated explicitly since it has in general been the case that each stream was comprised of a complete high performance ASR system in itself [12][14][19].

In the following sections some psycho-acoustic and phonetic motivations for introducing more heterogeneous signal processing in ASR is presented together with the relevant theory. Section 3 describes the databases and systems, in Section 4 we present the experimental results and Section 5 contains the conclusions and directions for future work.

2. HETEROGENEOUS INFORMATION IN MULTI-STREAM ASR

The focus of the work presented in this paper is on investigating the effects of introducing heterogeneous information in ASR, specifically in the multi-stream framework. Apart from the above mentioned pattern recognition issues, also several psycho-acoustic and phonetic studies have contributed to our motivation to initiate this work.

2.1. Psycho-acoustic and phonetic motivations

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 [8]. Studies that indicate specific processing in different frequency bands are also widely reported on. Experiments on the intelligibility of word pairs have shown that different phonetic features are transmitted in different temporal-frequency slots [7]. A related conclusion is made in [17] where it was shown that the optimal frequency range for recognising a phoneme in restricted transmission conditions is very dependent on the phoneme.

In line with this we choose to experiment with two ASR architectures: a general **multi-stream** system, where each stream processes the full spectrum, and a **multi-band** type of system, where each stream handles a limited frequency range. The questions we consider are: Can the performance and noise robustness of a multi-stream/multi-band recogniser be increased by introducing heterogeneous processing and in which way is the complementary information best utilised?

2.2. Theory

The introduction of heterogeneous information in a multi-stream system does not change the general formalism. The overall object of any speech recognition system is to find the most likely sequence of words $\mathcal{S}$ in a language $\mathcal{L}$ given a set of observed encoded features, $\mathbf{X}$. Within the HMM theory this breaks down to maximising the posterior probability:

$$P(q_j|\mathbf{x}_1,\ldots,\mathbf{x}_K) = \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_K|q_j)P(q_j)}{p(\mathbf{x}_1,\ldots,\mathbf{x}_K)} \quad (1)$$

where q_j is the jth class, K is the number of streams and $\mathbf{x}_k$ is the feature vector from the kth stream. In heterogeneous systems the feature vectors, $\{\mathbf{x}_1,\ldots,\mathbf{x}_K\}$ may be extracted with diverse signal processing methods. Multiple classifier theories view the estimation of $p(\mathbf{x}_1,\ldots,\mathbf{x}_K|q_j)$ as the problem of finding a fusion function, $f(\cdot)$ that combines the local stream likelihoods, $p(\mathbf{x}_k|q_j)$ to a common local likelihood:

$$p(\mathbf{x}|q_j) = f\left(W, \{p(\mathbf{x}_k|q_j^k), \forall k\}\right) \quad (2)$$

where W is a global set of weighting parameters. This approach is adopted for most multi-stream solutions too, and here we will use two widely used implementations of the fusion function: the sum and product rules, the geometric and arithmetic means respectively:

$$\text{Sum rule}: \quad p(\mathbf{x}|q_j) = \sum_{k=1}^{K} p(\mathbf{x}_k|q_j)^{\left(\frac{1}{w_k}\right)} \quad (3)$$

$$\text{Product rule}: \quad p(\mathbf{x}|q_j) = \prod_{k=1}^{K} w_k \cdot p(\mathbf{x}_k|q_j) \quad (4)$$

where $\sum_{k=1}^{K} w_k = 1$. In the experiments reported here we choose to use the simplest set of weights: namely equal and constant weights.

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 [16]. 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 [18] are added per utterance at SNR levels of 0, 6, 12 or 18dB. *Car noise*, *factory noise* and *Lynx helicopter noise* are chosen for their different spectral characteristics.

Three different feature processing methods are used for extracting basic features plus the energy: Mel frequency cepstrum coefficients (mfcc) [5], Perceptual linear prediction coefficients (plpc) [10] and J-rasta filtered plpc's (j-rasta-plpc) [11]. 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 hybrid MLP/HMM entities, are used. All MLP's 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 **full-band** system uses 12 basic features yielding a 39 dimensional feature vector. The MLP has 351 (9×39) input units and 1500 hidden units.
- The **multi-band** 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 MLP's have 162 (9×18), 162, 108 (9×12) and 108 input units and 1000, 1000, 660 and 660 hidden units per band respectively.

The baseline full-band and multi-band systems have a comparable number of parameters.

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

4. RESULTS

4.1. Heterogeneity in multi-band systems

As mentioned above, there are strong psycho-acoustic and phonetic motivations to investigate the effect of processing frequency regions differently. In this set of experiments we investigate whether clean speech performance can be increased by using heterogeneous feature extraction methods in a **multi-band** system.

In order to find out which feature type each stream could be based on, each sub-band classifier system was tested in isolation on clean speech. The purpose was to find out whether, for each band, there is a significant[2] difference in performance level among the different feature types. Such a difference is expectable since the feature extraction methods are based on different principles. The plpc's make use of an all-pole model of the spectrum and a Bark frequency scale and also try to account for some perceptual effects. The mfcc's are not based on any model assumptions and use a Mel frequency scale. Furthermore the additional j-rasta filtering is known mainly to increase the noise robustness.

The Word Error Rates (WER's) are shown in Table 1, and except for the first band there clearly is a difference in performance over the different feature types. For band

	j-rasta-plpc	plpc	mfcc
Band 1	35.12	35.50	35.78
Band 2	**31.50**	**29.64**	35.25
Band 3	39.06	**37.30**	43.70
Band 4	56.06	**52.72**	55.18

Table 1: Per band WER for multi-band baseline system. For each band the bold values identify which feature types significantly outperform the other feature types.

2 j-rasta-plpc's and plpc's outperform the mfcc's and in bands 3 and 4 the plpc's 'win'.

The findings are now used to test all possible combinations of these 'winning' feature types, and the WER's for the resulting heterogeneous multi-band systems are shown in Table 2. Two sets of experiments are carried out: using either 1) the sum or 2) the product rule of combination, eq. (3) and (4) respectively. Each table entry represents a different combination of features. The three bottom rows are the results from the homogeneous single feature type systems. The best heterogeneous multi-band systems significantly outperform the homogeneous systems for both rules of combination. A test of all 81 (3^4) possible combinations of feature-streams shows that the overall best combination is p+j+j+m resulting in WER's of 17.90% and 11.56% using sum and product rules respectively. None of the feature-streams in this superior combination obtain the highest WER in Table 1 which indicates that it is not just the performance of the streams but also the error complementariness which determine the overall performance of a multi-band system.

[2]at a 95% significance level.

	b1	b2	b3	b4	sum	prod.
Heterogeneous systems	j	j	p	p	19.23	**12.16**
	j	p	p	p	19.14	12.25
	p	j	p	p	**18.76**	12.36
	p	p	p	p	19.34	13.32
	m	j	p	p	19.81	13.40
	m	p	p	p	19.81	13.40
Homogeneous systems	p	p	p	p	19.34	13.32
	j	j	j	j	19.44	13.19
	m	m	m	m	21.73	15.09

Table 2: WER for different feature-stream combinations. The bold values indicate the best performing systems for each combination rule. j=j-rasta-plpc,p=plpc,m=mfcc.

4.2. Heterogeneity in multi-stream systems

Encouraged by the multi-band experiments we proceed to employ heterogeneous information in a **multi-stream** recogniser. The system is comprised of either two or three of the full-band recognisers, thereby having two or three times as many parameters as the baseline full-band systems. To equalise this, two other system configurations are tested: first a homogeneous full-band system boosted with two or three times as many hidden units (3000 or 4500) as the baseline full-band system, and second, another type of heterogeneous system trained on feature vectors that are a concatenation of the ordinary j-rasta-plpc, plpc and mfcc feature vectors. The MLP used in this system has 1053 ($9 \times 39 \times 3$) inputs, 1500 hidden units and 33 outputs, thereby tripling the number of parameters.

The results from testing all systems on clean speech are shown in Table 3. The multi-stream systems are tested using either the sum or the product rule for recombination. As was also observed with the multi-band systems, the product rule outperforms the sum rule in general. However, for the multi-stream systems a much smaller degradation in performance when using the sum rule instead of the product rule is observed. The sum rule tends to be a more severe way of fusing when the individual streams have a low performance [13], which is the case for the multi-band streams as seen in Table 1. The observations made in the following are drawn from results obtained using the product rule.

The heterogeneous multi-stream systems have an equal or lower WER than the homogeneous full-band systems to which they are comparable; however, the performance in-

Heterogeneous	Features	Sum / Prod. rule
2 full-bands	j + p	6.27 / 5.67
	j + m	6.70 / 5.97
	p + m	7.73 / 7.41
3 full-bands	j + p + m	6.72 / 5.80
concat. vectors	j / p / m	5.91
Homogeneous	Features	1500/3000/4500 HU
1 full-band	j	7.26 / 7.09 / 6.55
	p	7.39 / 7.41 / 7.32
	m	8.22 / 7.90 / 7.64

Table 3: WER's for heterogeneous systems and comparable homogeneous systems. 'HU' = hidden units.

crease is only significant for the j+p and j+m combinations. The j+p configuration even outperforms the very best homogeneous system, despite having only two-thirds the number of parameters. In contrary the p+m can not improve the performance of the corresponding homogeneous full-bands based on 3000 hidden units. It is clear that the choice of which features to combine is crucial, and further research is needed to fully understand in which way redundancy and diversity in features are best exploited.

The results in Table 3 also enables to compare two different systems based on heterogeneous processing. Both systems reduce WER in comparison to their homogeneous counterparts. However, it is not possible to conclude which heterogeneous system is better: the multi-stream system or the system using concatenated feature vectors. Arguably the multi-stream system is far more restricted in its abilities to exploit any correlations between the features. On the other hand, even though the two systems employ a comparable number of parameters, the MLP in the system with the concatenated feature vectors requires more training data since the input vector has a higher dimensionality than the vector for the multi-stream MLP's.

To test whether the introduction of heterogeneous information affords an increase in *noise robustness*, all the above mentioned systems are tested with added Factory, Car og Lynx helicopter noise (full details can be found in [4]). We obtained comparable results for the homogeneous and heterogeneous systems. Based on these experiments we are unable to conclude whether introducing heterogeneous information significantly increases the noise robustness. The j-rasta-plpc's are themselves very noise robust and adding the extra plpc and mfcc streams do not further increase the noise robustness. One explanation might be that j-rasta-plpc's are more noise robust for all phonemes. To test this we analysed the frame errors and found that this is not the case, and that there is a significant *difference* in performance between the ability of the features to classify the different phonemes. This suggests, that the combination methods employed in these experiments are not capable of fully exploiting the advantages of the different feature extraction methods, even though performance diversity does exist and can be observed at the frame level.

5. CONCLUSIONS

It has been shown that it is possible to increase clean speech performance in multi-stream and multi-band systems by increasing the heterogeneity of the signal processing employed in the systems and without increasing the number of system parameters.

The experiments reported here have employed three rather standard feature extraction techniques but results have encouraged us to continue exploring the potential advantages of more specifically designed stream processing. With respect to noise robustness the experiments have also made it clear that there is a need for more intelligent fusion of the information streams; this should involve taking into account advantages and weaknesses of the different streams. It becomes a particular interesting problem if streams are more specifically designed to specialise, for example on different phonetic segments, different acoustic-phonetic features, different gender etc.

6. ACKNOWLEDGMENTS

This material is based upon research carried out partly at the Dalle Molle Institute for Perceptive Artificial Intelligence, Martigny, Valais, Switzerland. We would like to thank Hervé Bourlard, Chafic Mokbel, Astrid Hagen, Hervé Glotin and Andrew Morris of IDIAP for valuable discussions.

7. REFERENCES

[1] H. Bourlard. Non-stationary multi-channel (multi-stream) processing towards robust and adaptive ASR. In *Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pages 1–10, Tampere, Finland, 1999.

[2] 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.

[3] C. Cerisara, J.-P. Haton, J.-F. Mari, and D. Fohr. A recombination model for multi-band speech recognition. In *Proceedings ICASSP-98*, pages II–717–II–720, Seattle, USA, May 1998.

[4] H. Christensen. Some experiments on the introduction of heterogeneous information into a multi-stream framework. Technical report, Center for PersonKommunikation, Aalborg University, 1999.

[5] 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.

[6] S. Dupont and H. Bourlard. Using multiple time scales in a multi-stream speech recognition system. In *Proc. Eurospeech '97*, pages 3–6, Rhodes, Greece, September 1997.

[7] 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.

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

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

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

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

[12] 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.

[13] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.

[14] N. Mirghafori. *A Multi-Band Approach to Automatic Speech Recognition*. PhD thesis, University of California at Berkeley, Berkeley, California, USA, December 1998.

[15] N. Nilsson. *Learning Machines*. McGraw-Hill, 1965.

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

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

[18] 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.

[19] S.-L. Wu. *Incorporating Information From Syllable-Length Time Scales into Automatic Speech Recognition*. PhD thesis, International Computer Science Institute, Berkeley, California, USA, May 1998.