SOME APPLICATIONS OF A PRIORI KNOWLEDGE IN MULTI-STREAM HMM AND HMM/ANN BASED ASR

Andrew Morris

Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)
P.O. Box 592, 4 Rue du Simplon, CH-1920, Martigny, Switzerland
morris@idiap.ch, http://www.idiap.ch/~morris

Abstract

Multi-band ASR was largely inspired by the extremely high level of redundancy in the spectral signal representation which can be inferred from Fletcher's product-of-errors rule for human speech perception. Indeed, the main aim of the multi-band approach is to exploit this redundancy in order to overcome the problem of data mismatch (while making no assumptions about noise type) by focusing recognition on sub-bands estimated to contain reliable, or "clean speech like", data.

However, multi-band processing also presents the opportunity to introduce a number of other ideas from phonetics, non-linear phonology and auditory processing into the recognition process. In particular: we can weight sub-bands, or sub-band combinations, according to the most likely frequency range of characteristic features for the phoneme whose presence we are testing for; we can allow some degree of asynchrony between sub-bands, and we can preprocess each sub-band according the kind of acoustic features which we expect to find there.

Besides combining sub-band experts, we can also combine multiple full-band experts, where each expert is perhaps suited to extracting complementary sources of speech information, or is robust to different kinds of noise. In this article we present an outline of some of the recent work at IDIAP, and cooperating institutions, in bringing together ideas from different areas of speech science within the framework of multi-stream HMM and HMM/ANN based ASR.

Keywords: noise robust ASR, multi-band and multi-stream processing, non-linear phonology, combination of experts

1. Introduction

Multi-band processing was developed primarily with a view to exploiting spectral redundancy for the purpose of robust speech recognition in noise [Morris, Hagen, Glotin & Bourlard, 1999]. However, the sub-band specific processing and asynchronous combination which fit naturally into the multi-band ASR framework also tie in directly with well established processing ideas from the areas of speech acoustics, and non-linear phonology [Mirghafori, 1999]. Furthermore, the methods used for combining sub-band experts in HMM/ANN hybrid ASR systems[1] can also be used as a simple means of combining multiple data streams of any kind, i.e. as many sources of complementary speech information as we like.

The modelling of evidence combination is a well established field, and before going into how multi-stream processing can be applied in ASR, in Section 2 we provide a brief introduction to the principles behind multi-stream processing in general. Then in Section 3 we discuss some of the evidence for multi-stream processing in human speech recognition. In Section 4 we discuss some of the evidence for multi-band processing in human speech recognition, and in Section 5 we introduce some recently developed models for multi-band and multi-stream processing in artificial speech recognition. In Section 6 we consider some of the different sources of speech information and some of the techniques by which they can be obtained. The article ends with a short discussion of the ideas presented, and a conclusion.

2. Multi-stream processing

Multi-expert systems arise in many different fields of data classification and function approximation in general. These systems have a number of proven theoretical and practical advantages, of which the following are of particular relevance to ASR:

- **Hierarchical systems of experts reduce problem perplexity:** Unsupervised-training can be used to train a hierarchical system of experts together with a gating network for expert selection. The gating network may be trained to use large scale features for expert selection, so training each expert on a subregion of the input data space, with correspondingly reduced perplexity [Jordan & Jacobs, 1994][Waterhouse & Robinson, 1994].
- **Linear combination of multiple experts can improve generalisation:** When expert outputs are linearly combined (even as a simple average), the expected committee error always decreases, both in theory [Bishop, 1995], and in practice [Raviv & Intrator, 1996]. This error will also decrease further if the spread of the experts' predictions can be increased without increasing the expected errors of the individual members. Different experts can be obtained by using different parametric functions and/or by varying the data used to train each expert, by different preprocessing or by adding different noise.

3. Multi-stream processing in human speech recognition

In any recognition process it is advantageous to constructively combine as many sources of information as are available [Morgan, Bourlard & Hermansky, 1998]. It is known that the

1. In HMM/ANN based ASR an ANN, typically an MLP, first transforms each acoustic feature vector into a vector of posterior phoneme probabilities. These posterior probabilities are then divided by their prior probabilities to provide scaled likelihoods, which are then used by the HMM for viterbi decoding

human auditory system is hardwired to combine visual with acoustic information, so that perceived phoneme category is directly influenced by lip movements [McGurk & McDonald, 1976]. ASR experiments have demonstrated that combining mouth shape with acoustic data can strongly improve recognition performance with noisy speech [Dupont & Luettin, 1998].

Further clear evidence for the use of multiple experts in the mammalian auditory system is seen in the cochlear nucleus, the first stage of central auditory processing. Each fibre in the auditory nerve splits and carries the same data through about seven different types of specialised nerve cell, each type having a very different characteristic response. The outputs from these cells are recombined at higher levels of processing [Pickles, 1988].

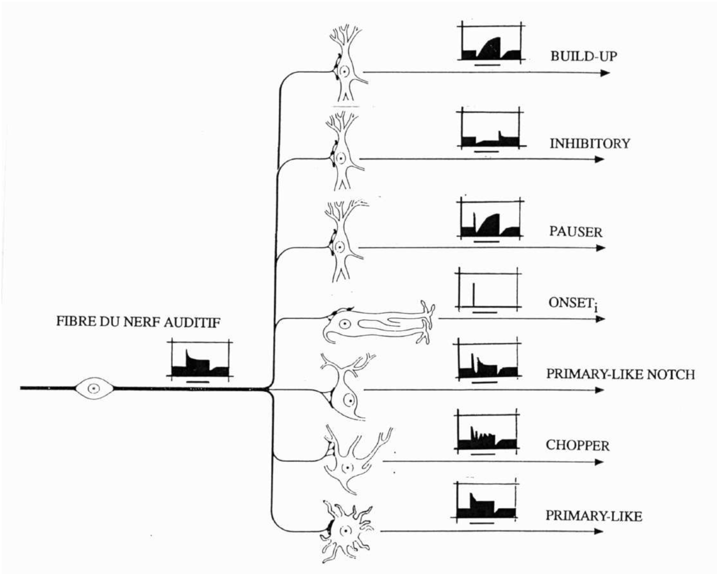


Figure 1. Different Cochlear Nucleus cell types and PSTH characteristics. Responses vary not only with cellular morphology, but also with the number and type of inputs from the auditory nerve or from other neurons in the cochlear nucleus [source?]

There are many possibilities in ASR for combining evidence from different data streams, such as vision with acoustics, or acoustic features from different time scales [Wu, Kingsbury, Morgan & Greenberg, 1998]. See also Table 1.

4. Multi-band processing in human speech recognition

While investigating the effects of band limited noise on human hearing, Fletcher [Fletcher, 1922] established a result, more recently publicised by [Allen, 1994], which is now commonly known as the "product-of-errors" rule (PoE rule), or Fletcher-Allen principle:

> *In human perception, the error rate for full-band perception is equal to the product of the sub-band error rates obtained through perception of each sub-band on its own.*

$$P(error) = \prod_i P(error_i) \tag{1}$$

Under the assumption of sub-band error independence, it follows from this rule that:

```
Full-band classification is correct if and only if
classification is correct in any sub-band
```

The product-of-errors rule therefore serves as proof of existence for a system which combines multiple guesses at the speech sound with an infallible mechanism for selecting the correct guess when it is present. This has strongly motivated the development of multi-band ASR.

5. *Multi-band processing in artificial speech recognition*

While the main motivation behind the multi-band approach is to exploit spectral data redundancy in a way which reflects the PoE rule for human speech perception, other potential advantages of the multi-band approach include:

- **Channel specific processing**: Different recognition strategies might ultimately be applied in each sub-band. For example, higher frequencies could use greater time resolution, and lower frequencies greater frequency resolution. It would also be possible to use sub-band specific speech subunits [Mirghafori, 1999].
- **Channel asynchrony**: Models discussed here use the same phoneme set for each sub-band expert, and force synchrony between experts, but it would be possible to permit some level of sub-band asynchrony [Bourlard & Dupont, 1996] (Figure 2).

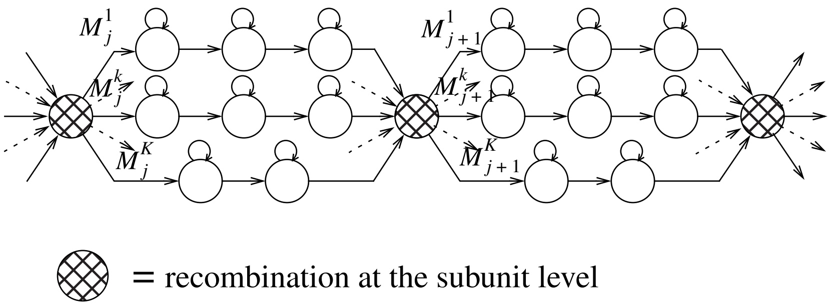


Figure 2. General form of a K-streams recogniser with anchor points between speech units (to force synchrony between different streams). Note that the model topology is not necessarily the same for the different subsystems.

However, when streams are not frame synchronous the complexity of the decoding algorithm required may be considerably greater than for a standard recogniser. Results to date have indicated that allowing asynchrony between streams does not give any significant performance improvement [Mirghafori, 1999].

The first multi-band ASR systems were based on the HMM/ANN (Hidden Markov Model/ Artificial Neural Network) model [Bourlard & Morgan, 1994]. In standard full-band ASR an MLP (Multi-layer Perceptron) is first used to transform the acoustic data into posterior phoneme probabilities[1], $P(q_k|x^n)$, for each word subunit, q_k, and data frame, x^n. Posterior probabilities from the MLP are then passed as scaled likelihoods into an HMM for decoding. In the early multi-band approach, one MLP is trained for each frequency sub-band, x_i, and the estimated posteriors $P(q_k|x_i^n;\Theta_i)$ from each MLP expert (here 4 experts, combined at the frame level) are then combined as either a weighted sum or product [Bourlard & Dupont, 1996][2]:

- *weighted sum, for posteriors combination:*

$$w_i = P(c_i|x)$$

$$P(q_k|x) \cong \sum_{i=1}^{d} w_i P(q_k|x_i;\Theta_i) \qquad \text{Identity when events } c_i \text{ mutually excl. and exhaustive.} \tag{2}$$

- *weighted product, for likelihoods combination:*

$$p(x|q_k) \cong \prod_{i=1}^{d} p(x_i|q_k;\Theta_i)^{w_i} \qquad \text{Identity when streams } x_i \text{ conditionally independent.} \tag{3}$$

Combined likelihoods from HMM experts, or combined posteriors as "scaled likelihoods" from ANN experts, are then passed to an HMM for decoding, as with the full-band system. If the data processed by each expert is both clean and independent of the data processed by other experts, then this approach is satisfactory. However, the data in spectral sub-bands is *not* independent, and independent sub-band processing cannot access joint spectral information, such as spectral envelope shape, which carries important information for phoneme discrimination. As a result the above combination rule does not perform competitively with clean speech. This problem can be overcome by extending this model to combine experts not just from each sub-band, but from every sub-band combination [Hermansky, Tibrewela & Pavel, 1996][Morris, Hagen, Glotin & Bourlard, 1999]:

- *full-combination weighted sum, for posteriors combination:*

$$P(q_k|x) \cong \sum_{i=1}^{2^d} P(c_i|x) P(q_k|x_{c_i};\Theta_i) \tag{4}$$

The "full combination" multi-band approach in Eq.4 gives a particularly strong advantage with narrow-band noise, while maintaining state-of-the-art performance in clean speech (see Section 7).

---

1. *See Nomenclature section for full definition of all mathematical symbols used.*

2. *Note that both a weighted product rule for posteriors combination and a weighted sum rule for likelihood combination can be obtained from Eqs. 2 and 3 respectively by direct application of Bayes' rule:* $p(x|q)P(q) = P(q|x)p(x)$.

6. Complementary sources of speech information

Some sources of information, such as harmonicity, synchrony and inter-aural time delays, are very important for signal-noise separation, but if the signal is clean then this information is not useful. Other sources of acoustic information which can be used for distinguishing different speech sounds exist in several forms and over a range of different time scales:

Scale 0.1-100 ms	Scale 10-1000 ms
short-term spectrum + differentials	amplitude modulation spectrum + differentials
abrupt energy transitions = pitch	abrupt energy transitions = phoneme transitions
voicing, glottalisation	phonological / articulatory constraints

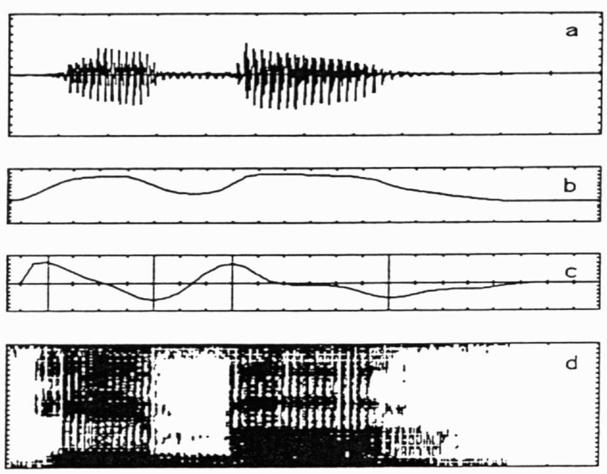
Table 1: Complementary sources of speech information at different time scales

Most common ASR systems use only the short-term spectrum and its time differentials, or some secondary features derived from these, such as MFCCs. One reason for this is that with clean speech no further information sources are necessary to achieve an acceptable level of recognition. Another reason is that it is perhaps not clear how all of these different kinds of information can be constructively combined. In the multi-stream ASR approach presented here, this combination is very straightforward.

We briefly describe below how some of these less standard features can be obtained.

6.1 Detecting phoneme transitions

Sub-band transitions are detected (Figure 3) using a simple model based on the function of onset detector cells found in the cochlear nucleus [Morris, 1992]. Phoneme transitions are then detected (Figure 4) by grouping sub-band transitions into onset or offset clusters.



a. waveform for a typical VCV utteranc
[eba]

b. smoothed spectral energy sum across
single frequency subband (band 1 of 4

c. On and Off positions detected at maxs
& mins in energy difference

d. the spectrogram for this signal

Figure 3. Onset and Offset transition detection in each sub-band [Morris, 1992]

```
1.\........../....../...../......\............/........\.........\......\........
2...\........./....../..../......................................\.....\.......
3...\................/..../........\...................\..........\.......\.....
4..\................./............\......../.........\.........\...........\..
5..\.................../..........\......../.........\.........\.......\......

6..\........../......../...........\........../........\.........\........\.....
7..kk.........kk.......ko..........om.........mU........Un........nn.......n-....
```

Figure 4. Phoneme transition detection: Lines 1 to 5 show onset and offset transitions detected for the Spanish word "comun" /KomUn/ from the test set. Line 6 shows detected transition cluster centres. Line 7 shows the estimated labelling [Morris & Pardo, 1995].

guess \ true	FV	VF	VR	RV
FV	90.2	0.0	0.0	0.3
VF	0.0	81.5	0.3	0.0
VR	0.0	0.0	54.5	0.0
RV	0.0	0.0	0.2	62.5

Table 2: Broad-class % confusion matrix using sub-band transition clusters in Spanish

F=fricative, V=vowel, R=r or rr. (cols do not sum to 100 as some data was unclassifiable)

Table 2 shows the confusion matrix for broad-class phoneme classification using sub-band transition clusters with a standard unsupervised clustering algorithm (the Kohonen map) (rr is the Spanish rolled r). This demonstrates that these clusters carry considerable information for phoneme discrimination.

6.2 Detecting glottalisation

In [Hagen, Shattuck-Hufnagel & Noeth, 1999], automatic detection of glottalisation (present vs. not present), using an MLP, achieved 64% frame recognition rate. Glottalisation is a

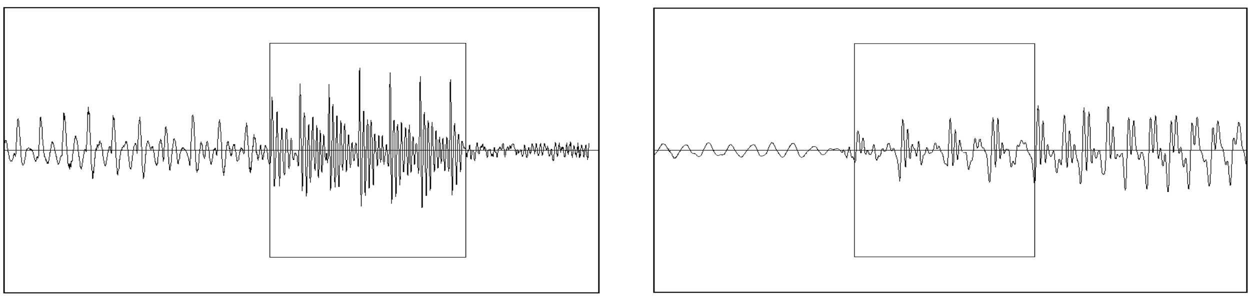


Figure 5. Two hand labelled examples of glottalisation from a study [Batliner, Burger, Johne & Kie, 1999] in which six different glottalisation contexts were identified.

distinctive acoustic feature which can be used to complement the usual set of phonemes.

6.3 Articulatory feature constraints

Various techniques have been developed to obtain articulatory parameters from speech acoustics. This problem is known as acoustic-articulatory inversion. If joint articulatory-acoustic data is available, then a parametric function, such as an MLP, can be trained to perform this inversion. The speech signal results from acoustic filtering by the vocal tract of a glottal excitation, so in the common case that articulatory data is not available, one approach is to apply some form of inverse-filtering [Schroetner & Sondhi, 1994]. Another is to *infer* the articulatory data by introducing latent variables into a Bayesian Network in a causal structure which suitably reflects the role of the articulators (or any other kind of explanatory variables for that matter) in speech production [Zweig, 1998][Conwell, Dawid, Lauritzen & Spiegelhalter, 1999] - although this would normally require two-pass recognition.

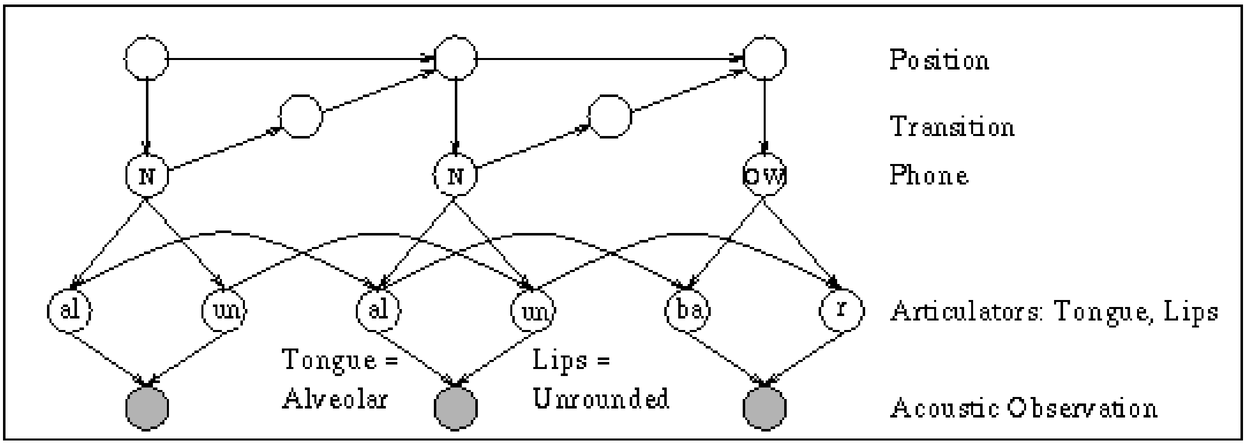


Figure 6. Articulatory parameters can be inferred from acoustics in a Bayesian Network.

Whichever way the articulatory parameters are obtained, they have not yet been widely accepted by the speech community as a reliable way of improving ASR performance. However, in the context of multi-stream combination, what is important is not whether these features are sufficient on their own for robust ASR, but whether they can *add* any information which is not already provided by the standard acoustic features.

6.4 ***Phonetic features constraints***

It has often been attempted to replace acoustic features with phonetic features [Stephenson, 1998][King, Stephenson, Isard, Taylor, & Strachan, 1998]. For each phonetic feature class (e.g. manner) an ANN expert is trained to classify acoustic data into probabilities for each phoneme feature sub-class (e.g. manner>nasal). In recognition the acoustic vector for each frame is replaced by a vector consisting of the concatenated vectors from each ANN expert

Table 3 shows the phonetic feature categories assigned to two of the phonemes in the TIMIT database. Table 4 shows the confusion matrix which results when an MLP is trained to distinguish sub-classes of the *manner* class.

	centrality, 4	continuant, 3	frontback, 4	manner, 6	phonation, 3	place, 14	roundness, 3	tense, 3
aa	central	cont	back	vowel	voiced	low	unrounded	tense
b	nil	non-cont	front	occlusive	voiced	labial	unrounded	non-tense

Table 3: Example phonetic features for two TIMIT phonemes.
(The number in "centrality, 4” etc. refers to the number of values each feature can take)

Manner	sil	approx	fric	nasal	occl	vowel
sil	89.2	1.4	2.0	1.0	3.2	3.1
approx	0.8	70.8	1.7	1.1	1.2	24.3
fric	2.1	1.1	87.7	0.8	5.0	3.3
nasal	2.1	3.1	2.3	79.9	3.3	9.2
occl	3.0	1.0	5.0	1.8	87.0	2.3
vow	0.5	5.2	1.2	0.8	0.9	91.4

Table 4: Typical feature state confusion matrix

As with articulatory features, recognition results based on phonetic features generally do not fulfil their expectations. This may be another source of information which is of more use in complementing standard ASR experts than it is standing on its own.

7. Some recognition results

Test results for some of these multi-stream and multi-band ASR methods are shown below.

Candidate Features	Abbreviation
Mel Frequency Cepstral Coefficients	mfcc
Perceptual Linear Prediction	plp
J-Rasta-PLP (plp with noise suppression)	jrplp
amplitude Modulation SpectroGram	msg
long-term TempoRAl PatternS	trap

Table 5: Candidate acoustic features used in MLP expert combination tests

Table 6 shows results from [Ellis, 2000b].

Stream components	Avg WER ratio % to baseline
best 1-expert: msg	60.7
best 2-expert: plp + msg	49.4
best 3-expert: plp + msg + trap	44.9

Table 6: (plp, msg, trap) features, with AURORA HMM baseline ASR system

Table 7 shows results from [Christensen, Lindberg & Andersen, 2000].

Stream components	clean WER
best 1-expert: plp	6.57
best 2-expert: plp + jrplp	6.17
best 3-expert: plp + jrplp + mfcc	6.02

Table 7: (plp, jrplp, mfcc) features, with STRUT hybrid baseline ASR system

Table 8 shows previously unpublished multi-band results recently obtained at IDIAP

ASR system	snr 0 dB WER	clean WER
full-band HMM/ANN hybrid ASR baseline	32.7	8.0
4 band full-comb multi-band hybrid ASR	13.5	9.3

Table 8: Multi-band ASR, noise in band 4 only, plp features, equal weights, vs. STRUT baseline

Of their nature, these results are rather scattered and can only be compared within each table. However, all baseline models here are state-of-the-art HMM or HMM/ANN systems, so any improvements over these represent new records in ASR performance for each system.

8. Discussion

We have described how the multi-band and multi-stream approaches have arisen through combining ideas from many fields, including linguistics, psychoacoustics, auditory physiology and information theory, in the search for noise robust methods in ASR. The way in which different sources of speech information are combined in these multi-stream models does not take any account of the nature of the information being combined and will therefore be suboptimal in many specific cases. However, the simplicity of the combination procedure allows us to focus instead on the importance of bringing together a set of maximally complementary sources of speech information.

Having acquired a number of information streams, the question arises of which sources should be concatenated and processed as a single stream, and which should be processed by separate experts before the speech category probabilities output from these experts are combined. A reasonable hypothesis [Ellis, 2000a] is that data streams should be concatenated if their data is highly dependent, but should otherwise be processed independently.

Another basic question concerns stream weighting. There are a large number of candidate procedures for stream weighting. For some types of data (e.g. acoustic or visual) it would seem reasonable to base weighting on a running SNR estimate. But another very simple approach, which is also adaptive and does not depend on the nature of the data, is to base the weight for each expert on the entropy (distribution flatness, or mean negative log probability) of the posterior probability distribution which it outputs.

9. Conclusion

Application of a priori knowledge in recognition often involves tailoring an existing system in some task specific way which is therefore inherently limited to one domain of application. The multi-stream HMM/ANN hybrid recognition paradigm provides us with a simple model which enables us to bring together and combine expert knowledge from any number of sources in a single framework. Much of the generality of this approach derives from the powerful non-linear modelling capability and discriminative training of MLPs. In the hybrid system this is combined with the proven time series modelling ability of HMMs. Both of the ANN and HMM paradigms are undergoing continual development and this can only improve the prospects for multi-stream combination.

10. Acknowledgements

This work was carried out in the framework of both the EC/OFES RESPITE (REcognition of Speech by Partial Information TEchniques) project, the EC/OFES SPHEAR (Speech, Hearing and Recognition) project. Thanks also to Heidi Christensen, Dan Ellis, Astrid Hagen, Sacha Krstulovic and Todd Stevenson who directly contributed material for this work.

11. Nomenclature

$P(x)$	probability of "event x" occurring
$p(x)$	probability density at x of a continuous value
$P(x;\Theta)$	function with parameters Θ used to estimate $P(x)$
q_k	speech unit whose presence is being estimated
$P(q_k\|x)$	probability that data x is from q_k
x, x^n	data window vector at time step n
d	number of spectral sub-bands
x_i	i^{th} sub-band of x, $i = 1\ldots d$
c_i	i^{th} sub-band combination, $i = 1\ldots 2^d$
x_{c_i}	i^{th} sub-band combination of x, $i = 1\ldots 2^d$
$P(c_i)$	probability that combination x_{c_i} is best (largest clean) subset of x, $i = 1\ldots 2^d$

References

Allen, J. B. (1994) “How do humans process and recognise speech?”, IEEE Trans. on Speech and Signal Processing, Vol.2, No.4, pp.567-576.

Batliner, A., Burger, S. & Johne, B. & Kie, A. (1999) “A Classification Scheme For Laryngealizations”, Proc. ESCA’93, pp.176-179.

Bishop, C. (1995) Neural Networks for Pattern Recognition, Clarendon Press, Oxford, pp.365-368.

Bourlard, H. & Dupont, S. (1996) “A new ASR approach based on independent processing and recombination of partial frequency bands”, Proc. ICSLP’96, pp. 422-425.

Bourlard, H. & Morgan, N. (1994) Connectionist speech recognition - a hybrid approach, Kluwer Academic Publishers.

Christensen, H., Lindberg, B. & Andersen, O. (2000) “Employing Heterogeneous Information in a Multi-Stream Framework”, Proc. ICASSP’00 (in press).

Conwell, R.G., Dawid, A.P., Lauritzen, L. & Spiegelhalter, D.J. (1999) Probabilistic Networks and Expert Systems, Springer.

Dupont, S. & Luettin, J. (1998) “Using the multi-stream approach for continuous audio-visual speech recognition: experiments on the M2VTS database”, Proc. ICSLP’98, pp. 1283-1286.

Ellis, D.P.W. (2000) “Stream combination before and/or after the acoustic model”, Proc. ICASSP’2000 (in press).

Ellis, D.P.W. (2000) “The ICSI RESPITE AURORA Multistream recognizer”, http://www.icsi.berkeley.edu/~dpwe/respite/multistream/

Fletcher, H. (1922) “The nature of speech and its interpretation”, J. Franklin Inst., 193(6), pp.729-747.

Hagen, A., Shattuck-Hufnagel, S. & Noeth, E. (1999) “A Study on Glottalizations and their Automatic Detection”, poster at ICPhS: Workshop on Non-Modal Vocal-Fold Vibration and Voice Quality.

Hermansky, H., Tibrewela, S. & Pavel, M. (1996) “Towards ASR on partially corrupted speech”, Proc ICSLP’96, pp. 462-465.

Jordan, M. I., & Jacobs, R. A. (1994) “Hierarchical mixtures of experts and the EM algorithm”, Neural Computation, 6, pp.181-214.

King, S., Stephenson, T., Isard, S. Taylor, P. & Strachan, A. (1998) “Speech recognition via phonetically featured syllables”, Proc. ICSLP’98.

McGurk, H. & McDonald, J. (1976) “Hearing lips and seeing voices”, Nature, No.264, pp.746-748.

Mirghafori, N., (1999) “A multi-band approach to automatic speech recognition”, PhD Dissertation, University of California at Berkeley, Dec. 1998. Reprinted as ICSI Technical Report, ICSI TR-99-04.

Morgan, N., Bourlard, H. & Hermansky, H. (1998) “Automatic speech recognition: an auditory perspective”, Research Report IDIAP-RR 98-17.

Morris, A.C. (1992) “An information-theoretical study of speech processing in the peripheral auditory system and cochlear nucleus: Application to the recognition of French voiced stop consonants", PhD thesis, INPG, France.

Morris, A. & Pardo, J. (1995) "Phoneme transition detection and broad classification using a simple model based on the function of onset detector cells found in the cochlear nucleus", Proc. Eurospeech'95, pp. 115-118.

Morris, A. C., Hagen, A., Glotin, H. & Bourlard, H. (1999) "Multi-stream adaptive evidence combination for noise robust ASR", IDIAP-RR 99-26, & Speech Communication (in press).

Pickles, J. O. (1988) An introduction to the physiology of hearing, Academic Press.

Raviv, Y., Intrator, N. (1996) "Bootstraping with noise: an effective regularisation technique", Connection Science, Special issue on Combining Estimators (8), pp.356-372.

Schroetner, J. & Sondhi, M.M. (1994) "Techniques for estimating vocal-tract shapes from the speech signal", IEEE Trans. on Speech and Audio Processing, 2(1 part II), pp.133-150.

Stephenson, Todd (1998) "Speech recognition using phonetically featured syllables", Masters Thesis, Univ. of Edinburgh, Centre for Cognitive Science. http://www.cstr.ed.ac.uk/projects/espresso/pubs/todd_dissertation.ps

Waterhouse, S. & Robinson, T. (1994) "Non-linear prediction of acoustic vectors using hierarchical mixtures of experts", in Neural Information Processing Systems 7, Morgan Kaufmann.

Wu, S.-L., Kingsbury, B. E. Morgan, N. & Greenberg, S. (1998) "Performance improvements through combining phone and syllable scale information in automatic speech recognition", Proc. ICASSP'98, pp.459-462.

Zweig, G. G. (1998) "Speech recognition with dynamic bayesian networks", PhD thesis, Computer Science, Univ. of California, Berkeley.