IDIAP RESEARCH REPORT

Recent advances in the multi-stream HMM/ANN hybrid approach to noise robust ASR

Astrid Hagen [a] Andrew Morris

IDIAP–RR 02-57

12/2002

révisé en 04/2004

à paraître dans
Computer, Speech and Language

Institut Dalle Molle
d'Intelligence Artificielle
Perceptive • CP 592 •
Martigny • Valais • Suisse

téléphone +41−27−721 77 11
télécopieur +41−27−721 77 12
adr.él. secretariat@idiap.ch
internet http://www.idiap.ch

[a] Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento em Lisboa (INESC-ID), Lisboa, Portugal; astrid.hagen@l2f.inesc.pt, http://www.l2f.inesc-id.pt, from 1/2004: astrid.hagen@daimlerchrysler.com

[b] Universität des Saarlandes, FR 4.7 Allgemeine Linguistik, Saarbrücken, Germany; morris@coli.uni-sb.de; http://www.coli.uni-sb.de

Rapport de recherche de l'IDIAP 02-57

Recent advances in the multi-stream HMM/ANN hybrid approach to noise robust ASR

Astrid Hagen Andrew Morris

12/2002

révisé en 04/2004

à paraître dans
Computer, Speech and Language

Résumé. In this article we review several successful extensions to the standard Hidden-Markov-Model/Artificial Neural Network (HMM/ANN) hybrid, which have recently made important contributions to the field of noise robust automatic speech recognition. The first extension to the standard hybrid was the "multi-band hybrid", in which a separate ANN is trained on each frequency subband, followed by some form of weighted combination of ANN state posterior probability outputs prior to decoding. However, due to the inaccurate assumption of subband independence, this system usually gives degraded performance, except in the case of narrow-band noise. All of the systems which we review overcome this independence assumption and give improved performance in noise, while also improving or not significantly degrading performance with clean speech. The "all-combinations multi-band" hybrid trains a separate ANN for each subband combination. This, however, typically requires a large number of ANNs. The "all-combinations multi-stream" hybrid trains an ANN expert for every combination of just a small number of complementary data streams. Multiple ANN posteriors combination using maximum a-posteriori (MAP) weighting gives rise to the further successful strategy of hypothesis level combination by MAP selection. An alternative strategy for exploiting the classification capacity of ANNs is the "tandem hybrid" approach in which one or more ANN classifiers are trained with multi-condition data to generate discriminative and noise robust features for input to a standard ASR system. The "multi-stream tandem hybrid" trains an ANN for a number of complementary feature streams, permitting multi-stream data fusion. The "narrow-band tandem hybrid" trains an ANN for a number of particularly narrow frequency subbands. This gives improved robustness to noises not seen during training. Of the systems presented, all of the multi-stream systems provide generic models for multi-modal data fusion. Test results for each system are presented and discussed.

Abbreviations

AAC	Approximate AC
AC	All-Combination (or “Full Combination”)
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
DCT	Discrete Cosine Transform
EM	Expectation Maximisation
FA	Feature Analysis
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HSR	Human Speech Recognition
LDA	Linear Discriminant Analysis
LVASR	Large Vocabulary ASR
MAP	Maximum A Posteriori
MB	Multi-Band
MCE	Minimum Classification Error
MD	Missing Data
MFCC	Mel Frequency Cepstral Coefficients
ML	Maximum Likelihood
MLP	Multi-Layer Perceptron
MS	Multi-Stream
MSG	Modulation SpectroGram
NB	Narrow-Band
NLDA	Non-Linear Discriminant Analysis
PCA	Principal Component Analysis
pdf	Probability Density Function
PLP	Perceptual Linear Prediction
RASTA	RelAtive SpecTrAl
RBF	Radial Basis Function
RNN	Recurrent Neural Network
SMD	Soft Missing Data
SNR	Signal to Noise Ratio
STD	Standard
SVM	Support Vector Machine
TP	Transition Probability
WER	Word Error Rate

Notation

T	number of data frames in utterance
K	total number of HMM states
q_k	state k (i.e. k^{th} of K HMM states in model)
q_t	state hypothesized at time step t
Q_t	HMM state sequence $(q_1, \ldots, q_t)$
Q	full HMM state sequence, Q_T
x_t	feature vector at time step t
x_i	coefficient i of x
x^i	feature vector subband or stream i of x
$x^{(i)}$	feature vector subband- or stream- combination (i)
$\|x^{(i)}\|$	number of subbands or streams in $x^{(i)}$
X_t	feature vector sequence $(x_1, \ldots, x_t)$
X	full feature vector sequence, X_T
$X^{(i)}$	feature vector sequence $(x_1^{(i)}, \ldots, x_T^{(i)})$
β	number of subbands or streams in x
$\mathcal{B}$	number of combinations of 0 or more streams, $= 2^{\beta}$
$b^{(i)}$	event that combination $x^{(i)}$ is informative and rest of x is not
$B^{(i)}$	event that combination $X^{(i)}$ is informative and rest of X is not
$p(x\|q)$	likelihood of state q given feature vector x (not of x given q)
$P(q\|x)$	posterior probability of state q given feature vector x
$P(q)$	prior probability of class (e.g. HMM state) q

1 Introduction

Most state of the art automatic speech recognition (ASR) systems are Hidden Markov Model (HMM) based, with state distributions modelled by Gaussian Mixture Models (GMMs). The performance of HMM/GMM systems is still improving, but it has a long way to go. Most of these improvements are in pre- and post-processing (feature extraction, pronunciation models, language models) and model adaptation (to channel, noise and speaker), rather than to changes in the central HMM/GMM modelling apparatus. HMMs are the preferred means for modelling time sequences, but an acknowledged weakness with the standard HMM/GMM in ASR is the assumption of independence[1] between consecutive data frames which are separated by only 10 ms, when strong dependence persists for 50-100 ms. Feature windows spanning more than one time frame do not improve HMM/GMM performance (Morris et al., 2000). The main interest in the HMM/ANN model (Bourlard and Morgan, 1994; Hochberg et al., 1995; Bourlard and Dupont, 1997), see Figure 1, is that Multi-Layer Perceptron (MLP) Artificial Neural Networks (ANNs) (Rumelhart et al., 1986; Ripley, 1996) are more able than GMMs to capture the dynamic information in extended feature windows, with frame level performance increasing beyond GMM performance as window size increases up to 90 ms. Both GMMs and ANNs append time difference features to static features, which somewhat increases the size of the context window (from 1 to 7 windows for GMMs and from 9 to 15 for MLPs), but difference features are just a fixed linear function of the context window and do not enable the GMM to perform the kind of non-linear processing of which MLPs are capable. It is true that through the use of time difference features, context dependent speech units, and language models, HMM/GMMs are still competitive with HMM/ANNs[2]. GMMs are also more highly developed (at present) than ANNs to noise and speaker adaptation. However, under many circumstances the propensity of ANNs to model class posteriors $P(q|x)$[3], while GMMs model class likelihoods $p(x|q)$, makes them more suitable than GMMs for multi expert combination[4].

1.1 HMM/ANN hybrid

In the HMM/ANN hybrid, which we refer to here as the "standard HMM/ANN hybrid" (see Figure 1), an ANN is trained to output a posterior probability for each model state. In decoding these probability mass outputs are converted to scaled likelihoods, as in (1), and used to directly replace the state likelihoods which are normally modelled by GMMs. However, there are many further ways in which the time sequence modelling power of HMMs can be combined with the superior ability of ANNs to capture speech dynamics.

$$\frac{p(x_t|q_k)}{p(x_t)} = \frac{P(q_k|x_t)}{P(q_k)} \tag{1}$$

For classification purposes the ANN used in HMM/ANN hybrids is usually an MLP with one hidden layer of sigmoid units and an output layer of softmax units (one per class). It is trained with labelled data to (most commonly) maximise the mutual information or "cross entropy" between input features and target output class posteriors. When HMM/ANNs are used with sub-word units such as phonemes, it is usual to restrict the size of the ANN by using just one ANN output per phoneme and to use the scaled likelihood from this output for all states of this phoneme. Furthermore, while in HMM/GMM systems the state transition probabilities (TPs) used in decoding are estimated as part of the Expectation-Maximization (EM) training procedure, in HMM/ANN systems this is not the case

[1] Throughout this article, when we say that random variables A and B are "independent" we really mean that they are conditionally independent with respect to the message (e.g. class labels) C they encode, i.e. $P(A,B|C) \cong P(A|C)P(B|C)$.

[2] HMM/ANNs also benefit from these context constraints, but to a lesser extent.

[3] Please refer to the Notation section where an exact definition is given for all symbols used throughout his article.

[4] This is mainly because, unlike state likelihoods, the state posteriors output by ANNs are independent of input data dimension.

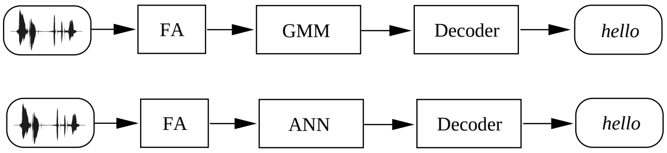

Fig. 1 – Standard HMM/GMM speech recognition system (top), with feature analysis (FA) followed by GMM state likelihood modelling and (Viterbi) decoder, and (bottom) a standard one-stream HMM/ANN system, with ANN state posteriors modelling.

and it is common practice to use the same fixed value (e.g. 0.5) for all TPs. However, while TPs contribute less than acoustic emission probabilities to decoder performance, HMM/ANN performance can be significantly improved (especially in noise) by better TPs estimation (Morris et al., 2002) and sometimes by using one ANN output per hidden state.

1.2 Multi-expert systems in ASR

The combination of multiple experts, where each expert has different error characteristics, provides a generic means of improving recognition robustness to unpredictable signal distortion. Recognition can be improved by combining multiple data models at one or more of the processing levels as follows (see also Figure 2).

1. **Feature combination :** concatenate data features from various sources. This is currently the most widely used form of feature combination.
2. **Posterior probabilities combination :** combine the K estimated posterior probabilities from each ANN into a single set of K probabilities.
3. **Hypothesis combination :** Performance can be further improved when multiple word sequence hypotheses from systems with different error characteristics are combined by combination schemes such as ROVER (Fiscus, 1997; Evermann and Woodland, 2000) or MAP AC MS (see Section 3.3) (Morris et al., 2001a).

In this paper we report on several extensions to the standard HMM/ANN hybrid which have recently made important contributions to the field of noise robust automatic speech recognition. The first extension to the standard hybrid was the (standard) “multi-band” (STD MB) hybrid (Bourlard et al., 1996; Hermansky et al., 1996), see Figure 3[5]. In this model the usual single ANN expert is replaced by a separate ANN trained on each frequency subband, followed by some form of posteriors combination, prior to decoding. However, while noise in any subband is thereby isolated, for all conditions except narrow-band noise this model results in significantly degraded performance.

MLP posteriors combination is usually by standard (weighted) sum or product (Hermansky et al., 1996; Tibrewala and Hermansky, 1997; Cerisara et al., 1998; Dupont, 2000; Kirchhoff et al., 2000), by voting (Halberstadt and Glass, 1998; Cerisara, 1999b), or by MLP (Bourlard and Dupont, 1996; Hermansky et al., 1996; Mirghafori, 1999). In the case of a weighted sum from N experts, if the errors from each expert are independent and unbiased, the expected square deviation from the target outputs will be reduced by a factor of N (Bishop, 1995). In practice the errors from different experts are always to some extent correlated, so that the variance reduction factor is smaller than N, but can still be large.

Analysis by the present authors has shown that the standard sum and product rules make a number of dubious assumptions. The STD sum rule makes the “one good subband” assumption that the data in just one subband is 100% reliable while the data in every other subband is completely

[5] Each ANN box in figures throughout this article also comprises any secondary feature processing, such as log energy scaling, Mel frequency scaling, PLP (Hermansky, 1990), concatenation, DCT, time differences, PCA, etc.

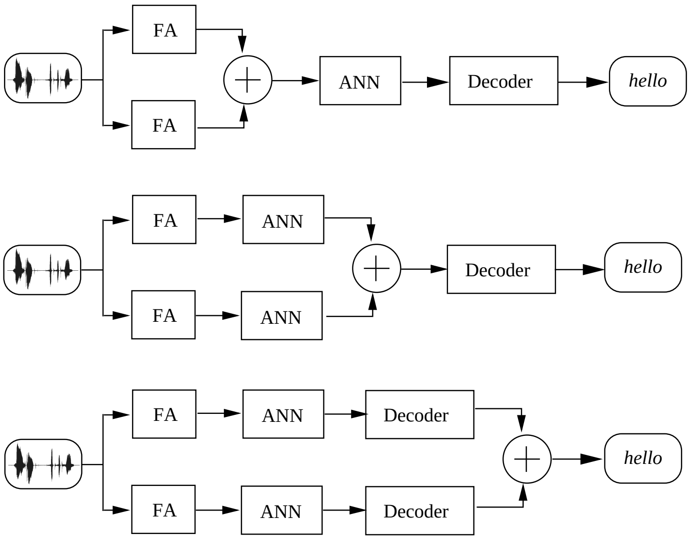

FIG. 2 – Three possible levels of multi-expert combination used by the HMM/ANN hybrid ASR systems reported in this artcile (top = feature level, middle = posterior proabilities level, bottom = hypothesis level).

uninformative. The STD product rule makes the "independence" assumption that each subband is (conditionally) independent. It also makes the "one good subband" assumption, but with suitable weight normalization this assumption can be avoided (see AAC sum rule, Section 3.1). STD MB MLP expert combination does not permit dynamic weighting and therefore assumes that subband relative reliability never changes, but this assumption is avoided in the "narrow-band tandem" model, Section 4.2.

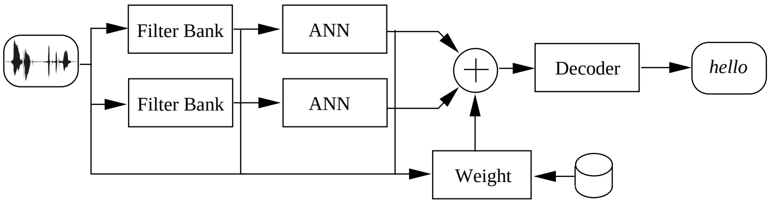

FIG. 3 – Standard multi-band HMM/ANN hybrid (STD MB), with two subbands. Combination is at the posteriors level only.

In the next section we discuss some observations concerning human speech recognition (HSR) which provided motivation for the original multi-band approach, because despite the poor performance of this early model, these "proof of existence" ideas from HSR may still provide valuable guidance in the future design of any kind of multi-expert ASR system.

2 Multi-band processing in human speech recognition

Most of the multi-expert techniques presented in this article arose from the standard multi-band model (Bourlard and Dupont, 1996; Hermansky et al., 1996; Goldberg and Riek, 2000), in which each frequency subband is processed by a separate acoustic model. This approach was motivated by Fletcher's work on human speech recognition (Fletcher, 1953), summarised in (Allen, 1994). Using nonsense consonant-vowel-consonant words and two frequency bands with varying cut-off frequency, Fletcher found that the human fullband error rate was related to the high- and low-pass error rates by the following simple formula, which we refer to as the "product of errors rule" :

$$\varepsilon = \varepsilon_1 \cdot \varepsilon_2 \tag{2}$$

where ϵ is the fullband error rate and ϵ_1 and ϵ_2 are the error rates from the lower and upper subband. This suggests that the two subband error rates are independent, and that a fullband error occurs exactly when there is an error in both high- and low-pass recognition. This is equivalent to saying that recognition is correct whenever either high- or low-pass recognition is correct. This means that human phone recognition is able to make at least two guesses at the phone identity, and to infallibly detect a correct guess when one arises. This rule identifies great potential benefit for any ASR system which could perform independent recognition in as many separate subbands as possible, providing

1. the advantage of having β guesses is not outweighed by the increased error rate for each guess
2. an infallible "oracle" can be found to identify which guesses are correct.

However, HSR tests since Fletcher's experiments, as well as ASR tests, now cast a strong doubt on the potential benefit of the STD MB model. HSR tests have now shown that

1. Fletcher's product of errors rule (2) does not hold in HSR with four or more subbands (Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985; Houtgast and Verhave, 1991; Steeneken and Houtgast, 1999).
2. the information carried by each combination of subbands (not just in neighbouring bands) is greater than the sum of the information carried in each subband taken alone (Grant and Braida, 1991; Steeneken and Houtgast, 1999; Lippmann, 1996; Silipo et al., 1999).

A combination of high- and low-frequencies (including a gap in frequency) often results in better recognition than a similar increase in band-width at low-frequencies by the use of a higher cut-off frequency (Lippmann, 1996; Silipo et al., 1999). Furthermore, in (Warren et al., 1995) it is shown that "sentences restricted to narrow spectral slits maintain a remarkably high intelligibility over an extended range of center frequencies, and that information contained in widely separated bands can be integrated to produce an increase in intelligibility that is much greater than simple additivity".

These findings in HSR are reflected by ASR tests which have shown that the early MB systems usually perform significantly worse than a fullband system in clean speech, even when perfect "oracle" expert selection is used (Hagen, 2001). The low performance of MB ASR in clean speech is due not so much to expert selection or weighting (which is relatively easy), but to the fact that the performance of every subband ANN expert is well below the performance of the fullband expert. This is not the case in HSR with just two subbands.

The hypothesis that HSR processes frequency bands independently is therefore no longer sustainable and in MB ASR we have to ensure that we at least model the joint processing of different subband combinations. In the next section we review the "all combinations" model which overcomes the independence assumption by training a separate classifier ANN on every combination of frequency subbands (or feature streams).

3 All-combinations multi-band and multi-stream systems

The first approach to overcoming the subband independence assumption was to pool fullband with subband experts (Bourlard and Dupont, 1997; Mirghafori and Morgan, 1998; Cerisara, 1999a; Mirghafori, 1999). However, while this combined system can prevent loss of performance with clean speech, it is not robust to wide-band noise and has no clear mathematical foundation.

3.1 All-combinations multi-band HMM/ANN hybrid

In the "all-combinations" multi-band HMM/ANN hybrid approach[6] (Hagen, 2001; Morris et al., 2001b) a separate ANN is trained for every combination of subbands, see Figure 4. This overcomes the inaccurate independence assumption, while retaining all of the potential advantages of a multi-band system.

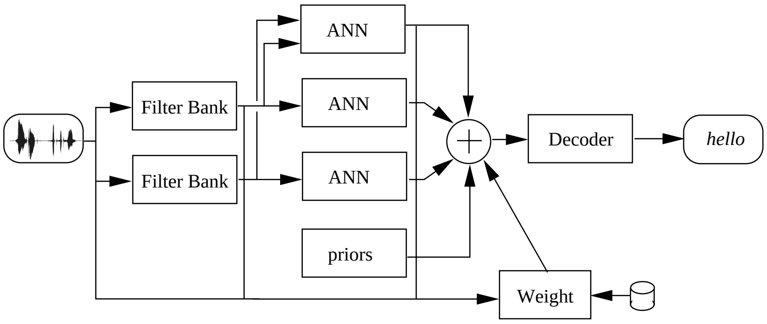

FIG. 4 – All-combinations multi-band (AC MB) HMM/ANN hybrid, with two subbands. An expert is trained for every possible combination of subbands. Combination is at both the feature and posteriors level.

The "AC sum rule" which is most commonly used for posteriors combination with the AC MB system having β subbands can be derived as follows. Let $x^{(i)}$ $(i = 1, \ldots, \mathcal{B})$[7] denote the vector of data coefficients from subband combination (i) for some given time frame x_t. Assume that each subband x^i is either 100% reliable or 100% uninformative (the "one good subband combination" assumption). Define $b^{(i)}$ as the event that every subband in $x^{(i)}$ is informative and all other data in x is uninformative. In this case the set of events $\{b^{(i)}\}$ are mutually exclusive and exhaustive. The AC sum rule can now be derived from first principles as follows.

$$\begin{aligned} P(q_k|x) &= P(q_k, \cup_i b^{(i)}|x) \quad (\{b^{(i)}\} \text{ exhaustive}) && (3)\\ &= \sum_{i=1}^{\mathcal{B}} P(q_k, b^{(i)}|x) \quad (\text{each } b^{(i)} \text{ mutually exclusive}) && (4)\\ &= \sum_{i=1}^{\mathcal{B}} P(b^{(i)}|x) P(q_k|b^{(i)}, x) && (5)\\ &= \sum_{i=1}^{\mathcal{B}} P(b^{(i)}|x) P(q_k|x^{(i)}) \quad (\text{definition of } b^{(i)}) && (6)\\ &= \sum_{i=1}^{\mathcal{B}} w_i P(q_k|x^{(i)}) && (7) \end{aligned}$$

Here the w_i are positive weights which sum to 1 and represent the probability that each event $b^{(i)}$ is true, given the data in x. Different approaches to how these weights can be estimated are discussed in (Hagen et al., 1999; Glotin and Berthommier, 2000; Heckmann et al., 2001).

The "one good subband combination" assumption made by the AC sum rule is an improvement on the "one good subband" assumption made by the standard MB sum rule, but it is still inaccurate. Not

[6] sometimes also known as the "full combination" approach

[7] $\mathcal{B} = 2^\beta$ is the total number of stream combinations having from 0 up to β subbands.

all coefficients in a subband need be equally informative and even noisy coefficients can sometimes impose strong constraints on the underlying clean data value, so they are not necessarily entirely uninformative.

With HMM/GMM systems, under certain conditions the subband combination posteriors can be derived from a single fullband model by marginalization[8] (Morris et al., 2000; Hagen, 2001), although the number of combination posteriors which is necessary to calculate can still be prohibitive if β is large. With the AC MB hybrid it is necessary to train 2^{β} separate ANN experts[9], so this system is limited to a small number of subbands.

However, if we assume stream independence, we obtain the following approximation to each subband combination posterior from the β one-band expert posteriors alone (Hagen et al., 2000; Morris et al., 2001b)

$$P_{ki} = P^{1-|x^{(i)}|}(q_k) \prod_{x_j \in x^{(i)}} P(q_k|x_j) \tag{8}$$

$$P(q_k|x^{(i)}) \cong P_{ki} / \sum_{j=1}^{\mathcal{B}} P_{ji} \tag{9}$$

This approximation, together with the AC sum rule (7), provides the approximated AC (AAC) sum rule. This rule can be used with the standard MB model, replacing the usual assumption that just one subband is reliable by the assumption that just one subband combination is reliable.

Results comparing the performance of the standard fullband, 4-subband standard MB and AC MB hybrid systems are presented in Figure 5[10]. It is of interest to note that the performance of the AAC sum rule is consistently better than that of the standard MB sum rule, given that the difference in computational complexity is negligible.

Tests in Figure 5 used the Numbers95 database (Cole et al., 1995) of continuously spoken digits. Noise conditions were clean (matched) and noisy (mismatch), with band-limited (stationary and siren) and wide-band car and factory noise from Noisex (Varga et al., 1992), each at 0 and 12 dB SNR (Hagen, 2001). Stationary band-limited Gaussian noise was added to one of the subbands at a time at 0 and 12 dB SNR. Throughout this article, all noise signals were artificially added to the sampled speech signal.

All of the results given in this section are for systems trained with clean speech. Word error rates (WERs) were calculated as average values over the different noise levels. All tests were run using both PLP and (more noise robust) J-RASTA-PLP (Hermansky et al., 1992) features. Feature sets consist of 12 raw features (together with frame energy), calculated on 25 ms windows with 12.5 ms shift from signals sampled at 8 kHz, plus first and second time difference features.

As can be seen in Figure 5, standard MB processing, employing either STD sum or STD product rules, improves performance only with narrow-band noises and non-robust features, while performance for both wide-band noise and clean speech is strongly degraded in comparison to the fullband HMM/ANN baseline. AC MB also gives an advantage with narrow-band noise for robust features, while not significantly degrading performance in clean speech. However, with wide-band noise and clean speech, none of these systems gives any significant improvement over the fullband baseline.

In the next section we look at the AC multi-stream (AC MS) approach. This uses the same architecture and combination rules as AC MB, but, unlike AC MB, it can improve over or equal the performance of the fullband baseline (using noise robust features) under all noise conditions.

[8] Closed form marginalization with respect to a subset of spectral coefficients is only possible with GMMs when features remain in the spectral domain (e.g. not possible if DCT has been applied to obtain cepstral coefficients).

[9] The number of ANN experts can be reduced to some extent by assigning a zero weight to certain subband combinations, such as all combinations with less than a given number of subbands.

[10] Combination weights w_i and state priors $P(q_k)$ used in tests throughout this article, unless otherwise stated, are all equal. In the authors' experience, the difference in performance due to different weighting methods is usually negligible compared to the difference due to different model architectures. State priors help if they are estimated accurately, but if one or more is inaccurate (e.g. due to few state occurrences), it is better to use equal priors.

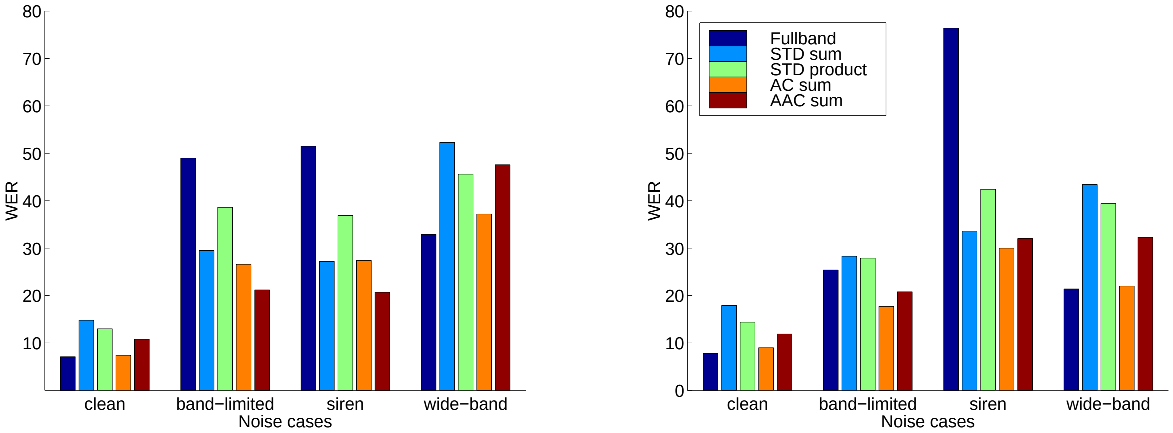

FIG. 5 – Test results (Numbers95 connected digits data) for the fullband baseline, the standard multi-band (STD MB) hybrid (with STD sum and STD product combination), the all-combinations multi-band (AC MB) hybrid (with AC sum combination), and the approximate all-combinations hybrid (with AAC sum combination), employing either PLP features (left) or J-RASTA-PLP features (right). All training was with clean speech only. WERs are given for clean speech, speech corrupted by stationary and dynamic artificial narrow-band noise, and average for wide-band car and factory noise. All noise results averaged over 0 and 12 dB SNR. Results from (Hagen, 2001).

3.2 All-combinations multi-stream hybrid

The AC MB system models noise more accurately than the standard MB system, but its performance is still limited due to the inaccurate assumption of all-or-nothing reliability for each subband. An alternative approach is to leave noise removal to noise robust features, or noise estimation and subtraction, and to apply the AC MB model to the combination of features from experts trained on multiple representations of the full speech signal. We call this the AC multi-stream (AC MS) approach (model as Figure 4, but with “Filter Bank” replaced by “Feature Analysis”).

The paradigm of using an ensemble of trained classifiers instead of a single classifier has been widely proposed in the literature (Hansen and Salamon, 1990; Jacobs et al., 1991; Jordan and Jacobs, 1994; Bishop, 1995). Ghitza (1994) showed that humans seem to use not only different frequency bands but also different time scales to capture short-term and long-term information simultaneously.

As mentioned in Section 1.2, the advantage of expert combination is greatest when each expert has different error characteristics and is unbiased. Expert diversity can be obtained by variation of one or more of the many factors involved in expert design, including : sensory mode, e.g. audio, visual (Tomlinson et al., 1996; Dupont and Luettin, 1998; Rogozan and Deléglise, 1998) ; training set ; training noise environment (Tumer and Ghosh, 1996; Shire, 2000) ; sample size ; analysis technique (e.g. MFCC (Davis and Mermelstein, 1980), PLP, RASTA) ; analysis time scale (Hagen, 2001; Hermansky et al., 2000; Ellis and Reyes-Gomez, 2001; Ghitza, 1994) ; window size ; derivative window size (Wu et al., 1998; Kirchhoff, 1998; Hagen et al., 2000) ; expert type (e.g. GMM, MLP, RNN) ; expert configuration (e.g. number of mixture components in GMM ; number and size of hidden layers in MLP) ; training objective (e.g. sum of square errors, cross-entropy, minimum classification error).

Experimental results in Figure 6 show that with clean speech the AC MS model leads to consistently (although not dramatically) improved results over the baseline hybrid using one stream of concatenated features. The three streams use state-of-the-art acoustic feature analysis techniques which are known to be powerful in rather diverse conditions and thus complement each other well. Although the RASTA feature stream when used alone performs significantly worse than the others, the additional use of each feature stream leads to a significant performance improvement.

In these tests the AC MS sum rule (with equal stream weights) was used to combine three different

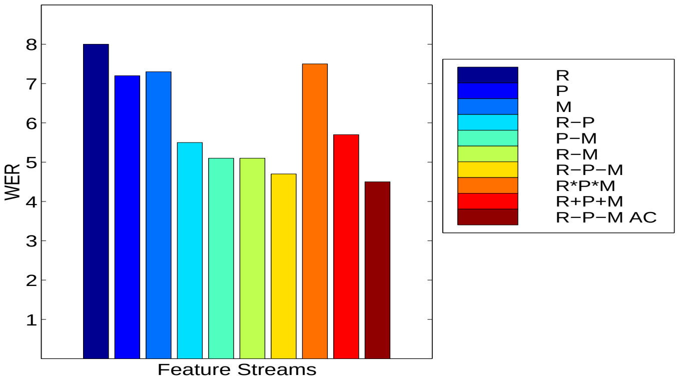

FIG. 6 – Test results (continuous speech from digits and numbers part of Portuguese SpeechDat corpus, clean speech) for the all-combinations MS HMM/ANN hybrid, employing PLP (P), RASTA (R) and MSG (M) features. Results are given for single streams, for feature concatenation (R-P,P-M,R-M,R-P-M) and for posteriors combination employing STD sum (R+P+M) and product (R*P*M) rules, and AC sum rule (R-P-M AC). Results are from (Hagen and Neto, 2003).

feature analysis techniques, and lead to improved results on clean speech, which was not possible with AC MB processing. In the next section we review a technique whereby the stream weights used by the AC MS sum rule are selected to maximise the a-posteriori utterance probability. This approach leads analytically to a simple method for combining any number of standard HMM/ANN or HMM/GMM utterance hypotheses.

3.3 Combination at hypothesis level

In maximum likelihood (ML) based adaptation a small number of parameters θ in a trained model with parameters (M, θ) are adapted during recognition of an utterance X to maximise the parameter likelihood, $p(X|M, \theta)$ and the utterance Q is then selected to maximise $p(Q|X, M, \theta)$. With posteriors based models, such as the AC MS HMM/ANN hybrid, with AC sum rule combination (6), it is not possible to evaluate the likelihood $p(X|M, \theta)$[11] (Hagen, 2001, 101–103), but it is possible to evaluate the posterior utterance probability $P(Q|X, M, \theta)$ for any given Q. In this section we review an approach by which combination weights for an AC MS hybrid system are adapted for each utterance to maximise the posterior utterance probability $P(Q|X, M, \theta)$ over all Q and all θ.

In (Morris et al., 2001a) two cases are considered. In the first, "static MAP weighting", the constraint is imposed that the same weighting must be used throughout each utterance. In this case we are effectively hypothesizing that each of the β streams of X are either 100% informative or uninformative throughout each whole utterance. There are therefore just 2^β possible informative subsets of X, as in the AC sum rule (7). In the second case, "dynamic MAP weighting", no such constraint is imposed. In this case the sum in (7) must run over all $2^{T\beta}$ possible subsets of X. In either case we can replace the frame based variables b, q, x, in (3) to (7) by the utterance based variables B, Q, X, to obtain the utterance based equivalent to (7) :

$$P_w(Q|X) = \sum_i w_i P(Q|X^{(i)}) \tag{10}$$

where $w_i = P(B^{(i)}|X)$. Equation (10) has form $A = \sum_i w_i a_i$ where a_i are fixed positive values for a given Q, $w_i \geq 0$ and $\sum_i w_i = 1$. It follows (see proof in Appendix A) that the weights which maximise

[11] State posterior probabilities $P(q|x)$ can be obtained from state likelihoods $p(x|q)$ using Bayes' rule, but likelihoods cannot be obtained from posteriors.

A are simply given by $w_i = 1$ for $i = \arg\max_j a_j$ and all other weights $= 0$[12]. This gives

$$\max_w P_w(Q|X) = \max_i P(Q|X^{(i)}) \tag{11}$$

The MAP solution with static weights can therefore be obtained by finding the MAP utterance Q^{MAP_i} for each expert i (making a note of $P_i = P(Q|X^{(i)})$), then selecting Q which has the maximum probability over all experts[13].

This procedure is formalised in (12) to (16).

$$\begin{aligned} Q^{MAP_i} &= \arg\max_Q P(Q|X^{(i)}) = \arg\max_Q P(Q)\frac{p(X^{(i)}|Q)}{p(X^{(i)})} && (12)\\ &\cong \arg\max_Q P(Q)\prod_t \frac{p(x_t^{(i)}|q_t)}{p(x_t^{(i)})} = \arg\max_Q P(Q)\prod_t \frac{P(q_t|x_t^{(i)})}{P(q_t)} && (13)\end{aligned}$$

$$P_i = P(Q^{MAP_i}|X^{(i)}) \tag{14}$$

$$j = \arg\max_i P_i \tag{15}$$

$$Q^{MAP} = Q^{MAP_j} \tag{16}$$

Static MAP AC MS weighting at the posteriors level therefore leads to a system where combination is effectively at the hypothesis level (see Figure 7). While it was previously well established that combination at hypothesis level can lead to improved recognition performance (Fiscus, 1997; Evermann and Woodland, 2000), static MAP weighting uses the Bayes optimal MAP objective (subject to the assumption of static weights), whereas most other hypothesis combination systems are based on intuitive ideas of alignment and voting.

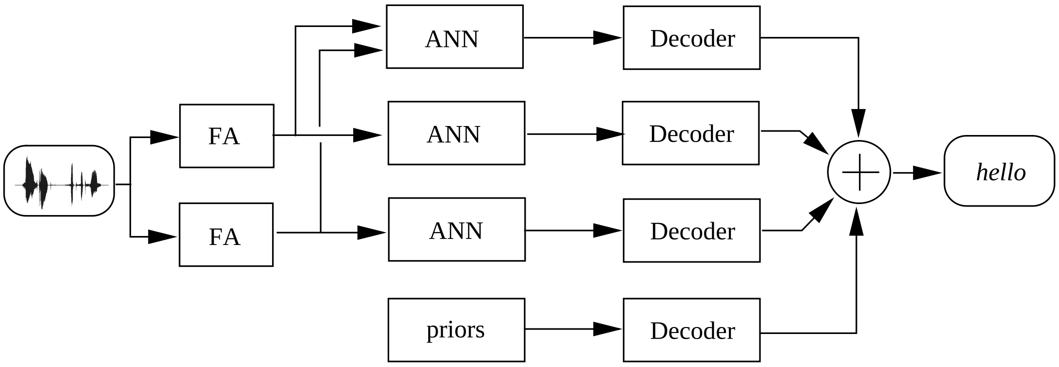

FIG. 7 – MAP AC MS combination model (two streams). An expert is trained for every combination of streams. With two subbands there are four possible combinations of zero or more subbands, as shown. For β subbands there are 2^β possible combinations. Although this method derives from "static MAP weighting" for posteriors level combination, in practice this is achieved by performing standard Viterbi decoding separately with each combination expert, and then selecting the hypothesis which has greatest MAP probability over all experts. Combination in the resulting system is at both feature and hypothesis level.

[12] The same MAP solution also results when weights are selected to maximise the weighted posteriors product (take logs, ignore weight normalization, and proceed with proof as for sum).

[13] $P(Q)$ below is modelled in the usual way as $P(q_1)\prod_{t=2}^{t=T} P(q_t|q_{t-1})$, not as $\prod_{t=1}^{t=T} P(q_t)$.

The MAP solution with dynamic weights can be found by selecting, during Viterbi decoding, at each time step t and for each q_k, the single combination expert which results in the largest posterior partial utterance probability, $P(Q_t|X_t)$. However, test results for dynamic MAP weights (not reported in (Morris et al., 2001a)) were not as good as for static MAP weights, probably due to an excess of free parameters, which resulted in over-fitting.

Due to the symmetry between likelihoods and posteriors in (13), static MAP AC MS combination can be implemented either by GMMs modelling $p(x_t^{(i)}|q_k)$, or by ANNs modelling $P(q_k|x_t^{(i)})$. In the case where GMMs can be marginalized over "missing streams" (i.e. where feature stream combination is by concatenation alone) static MAP AC MS decoding can be achieved through marginalization of the standard full-stream GMM (see Figure 8).

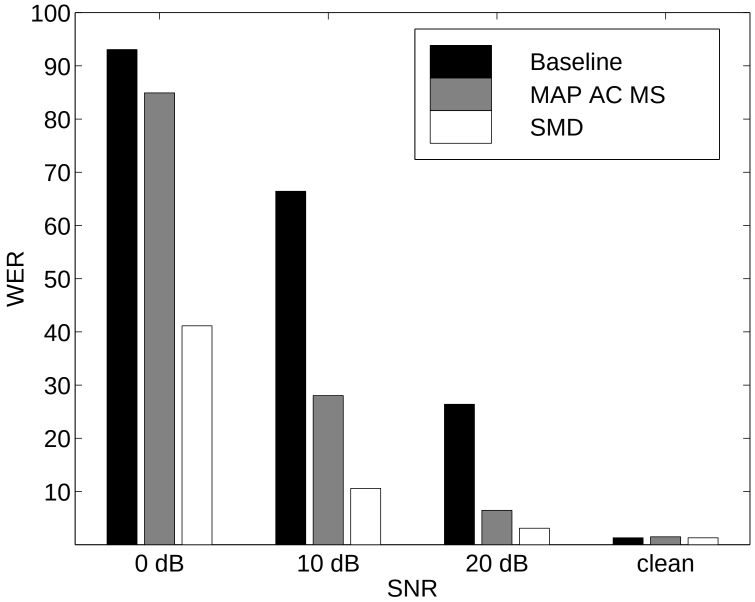

FIG. 8 – Results are compared for three systems using the same standard HMM/GMM models, trained on clean speech, but different decoding methods : Baseline (standard Viterbi) ; AC MS with "static MAP weighting" (Viterbi for each stream combination, with marginalisation over missing streams, followed by MAP hypothesis selection) (Morris et al., 2001a) ; SMD (Viterbi with marginalisation over estimated "missing data mask") (Barker et al., 2000). Test data is Aurora connected digits, with noises from test set A (Hirsch and Pearce, 2000). Training is with clean data. Feature streams are Mel spectral features and their first time differences[15].

Figure 8 compares baseline HMM/GMM, "Static MAP AC MS" and "Soft Missing Data" (SMD) recognition results, for an example where each system uses the same HMM/GMM model, (trained on clean spectral coefficients and their first time differences), but a different variety of Viterbi decoding. Here MAP AC MS performs much better than the baseline, although not as well as the SMD model. However, the AC MS system here has the triple handicap that (i) it is intended for combining clean data streams (noise should be removed beforehand, which it is not here), (ii) when ANN experts are used, rather than the GMMs which are used here, feature combination is not limited to simple concatenation (as it is here, in order to illustrate the important point that MAP AC MS can also be used with GMMs), and (iii) missing-data methods base data reliability on local SNR estimation, so cannot be used with multi-condition training (SNR level would no longer be a good indicator of data mismatch), while AC MS would work better with multi-condition training.

In this section we have seen how the "all-combinations" approach of training a classifier ANN for every possible stream combination, followed by posteriors combination, provides one way of combining multiple data streams without the assumption that each stream is independent, or that just one stream is reliable. However, this approach is only well suited to situations where the number of streams to be combined is not much greater than three, and stream reliability is all-or-nothing rather than graded. In the next section we review two recent "tandem" HMM/ANN ASR systems which also exploit the classification capacity of ANNs and avoid the assumption of stream independence, while training only one classifier ANN per subband or stream.

4 Tandem HMM/ANN systems

In "tandem" HMM/ANN ASR systems only one classifier ANN is trained per subband or stream. Training is with multi-condition data and instead of interpreting the ANN outputs as class posterior probabilities, they are exploited as discriminative and noise robust "non-linear discriminant analysis" (NLDA) features[16] (Fontaine et al., 1997) for input to a standard HMM/GMM or HMM/ANN system. The two tandem systems presented in this section have performed well in noise robust ASR tests. They are really nothing more than special forms of enhanced feature processing, and are therefore complementary to the above "all-combinations" multi-stream systems, which exploit the different advantages of expert combination.

With both of these tandem systems, ANN classifier outputs are always in the interval [0,1] and tend to be close to 0 or 1. This data therefore has a highly skewed bimodal distribution and is not well modelled by either ANNs or GMMs. More evenly distributed discriminative feature data with a greater dynamic range can be obtained either simply by omitting the final squashing non-linearity in the trained ANN, or by training the network with an extra hidden layer, and then taking the output from this hidden layer as discriminative features, instead of from the output layer.

Two forms of tandem HMM/ANN system are described below.

4.1 Multi-stream tandem HMM/ANN

The "multi-stream tandem hybrid" (MS-tandem) (Ellis and Reyes-Gomez, 2001) trains a separate ANN for a number of complementary feature streams, using multi-condition data. Principal component analysis[17] (PCA) is then used to orthogonalize the concatenated pre-squashed ANN outputs, which are then used as features for input to a standard HMM/GMM (see Figure 9).

Two MS-tandem systems were tested on the Aurora 2.0 task of connected digit recognition in noise (Hirsch and Pearce, 2000), test-A (matched noise types in training and testing) : (i) a single stream tandem employing 13 PLP features (together with their first and second time difference features) as input to the MLP-ANN (which uses 9 frames of context, 480 hidden units, and 24 output nodes), results in Figure 10. (ii) a two stream tandem system (13 PLP features and 28 MSG features (Greenberg and Kingsbury, 1997), MLP details as in (i)), results in Figures 10 and 11. The baseline system for all of these experiments constituted a standard HMM/GMM employing GMMs trained on multi-condition PLP features directly.

In Figure 10 we see that, in all noise conditions, the MS-tandem system gives improved recognition over the single stream tandem, and the single stream tandem improves over the HMM/GMM baseline (using the same features).

The MS-tandem system is very well suited to the Aurora test-A (matched noise conditions), see Figure 11. However, the advantage of this approach has been found to be much reduced both when different noise types are encountered in testing, and in large vocabulary ASR (LVASR) (Hermansky et al., 2000). The "narrow-band" tandem in the next section is more robust to mismatched noise conditions.

[16] i.e. as a non-linear generalisation of linear discriminant analysis (LDA) (Duda and Hart, 1973; Fukunaga, 1990; Haeb-Umbach and Ney, 1992), which is commonly used for feature data enhancement.

[17] All PCA features are retained, as performance fell with reduction in data dimension.

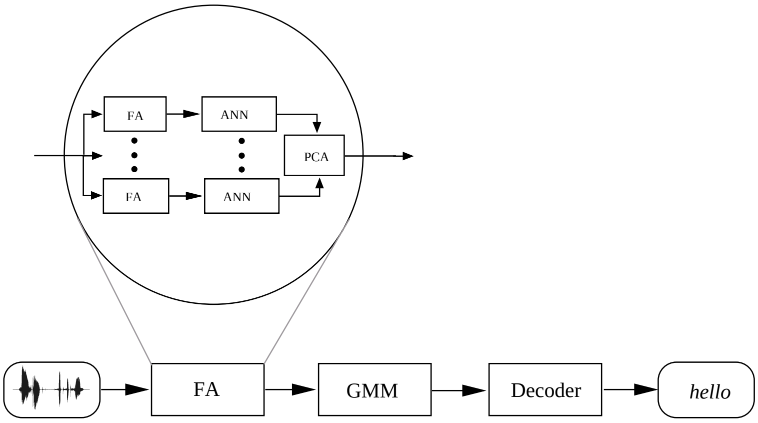

Fig. 9 – Multi-stream tandem (HMM/ANN) hybrid (MS-tandem). This can be regarded as a standard HMM/GMM system in which the FA module comprises several parallel and complementary feature analysers, each post-processed by an ANN before being concatenated and subjected to PCA orthogonalisation.

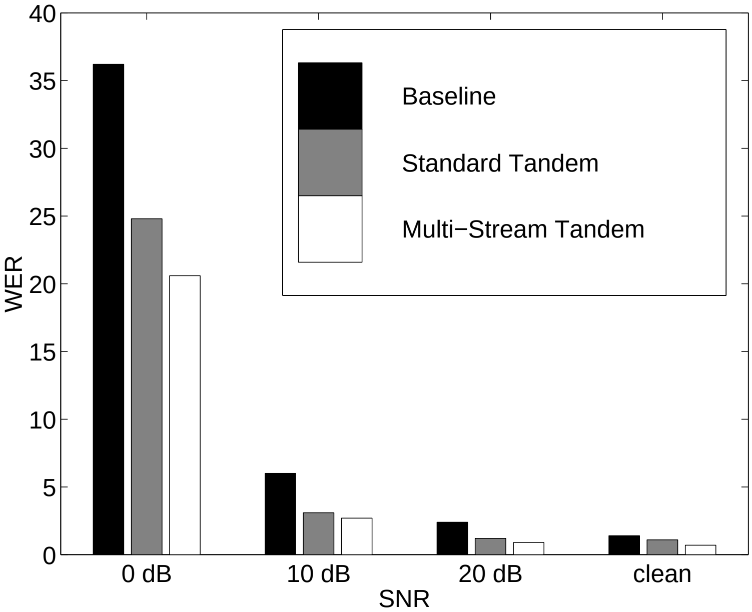

Fig. 10 – Test results (Aurora 2.0 connected digits data, test A = matched noise) for the HMM/GMM baseline, one-stream tandem (standard tandem) and multi-stream tandem HMM/ANN, under varying noise levels (WER scores are averaged over the 4 noise types). The first two systems employ PLP features. The multi-stream system uses PLP and MSG features. All systems were trained on multi-condition speech data. Results are from (Ellis, 2002).

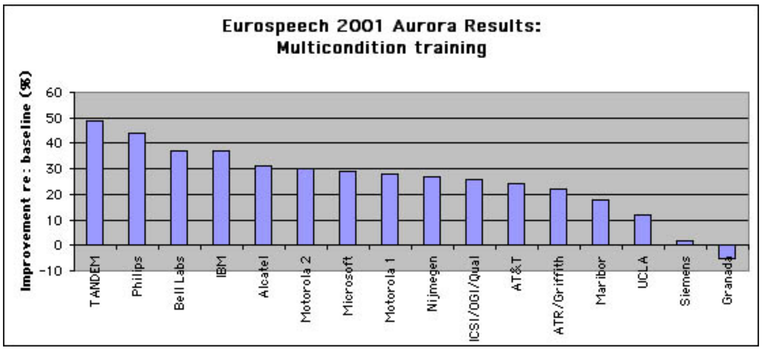

FIG. 11 – Test results comparing performance of systems from foremost speech research labs (Aurora 2.0 connected digits data, test A = matched noise). Figures are averaged over all noise conditions (MS-tandem system on left). Results from (Ellis and Reyes-Gomez, 2001).

4.2 Narrow-band tandem HMM/ANN

The "narrow-band tandem hybrid" (NB-tandem) system (Dupont and Ris, 2001), applies a separate NLDA preprocessing to each of a number of narrow frequency subbands. Each narrow-band ANN is trained using a small number of Bark scaled spectral features for clean data to which varying amounts of white noise have been added. This system operates on the principle that as the frequency range seen by each ANN is very narrow, the spectral shape of the noise is not detectable, so the type of noise encountered during recognition should not make any difference. In this system each narrow-band ANN was given an extra small hidden layer and NLDA features were taken as the output from this layer. The concatenated NLDA output from these trained narrow-band ANNs was input as features to a standard HMM/ANN hybrid (see Figure 12).

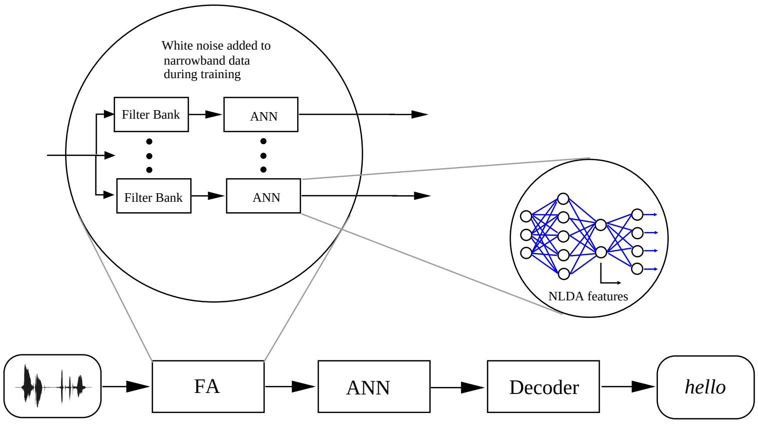

FIG. 12 – Narrow-band tandem (NB-tandem) HMM/ANN hybrid. This can be regarded as a standard HMM/ANN system in which the FA module comprises several parallel subband feature analysers which first extract standard features and then post-process these using ANNs which were trained on data corrupted with white-noise at different SNRs. FA outputs are taken from the second hidden layer of these ANNs.

Tests were made with the Aurora 2.0 connected digits database[18]. Two configurations of the NB-tandem system were set up : (i) a parameter-heavier version, with the multi-band MLPs having 1000 units in the first and 30 units in the second hidden layer. The combining MLP has 127 HMM states and uses 1000 units in its hidden layer and 3 frames of context. (ii) a lighter version employing approximately the same number of parameters as the baseline system by only employing 150 units in the first hidden layer of the multi-band MLPs, and only 500 hidden units in the first layer of the multi-band MLPs, only 500 hidden units and just one frame of input for the combining MLP. The baseline HMM/ANN was trained on clean data. Both baseline and NB-tandem systems used J-RASTA-PLP features.

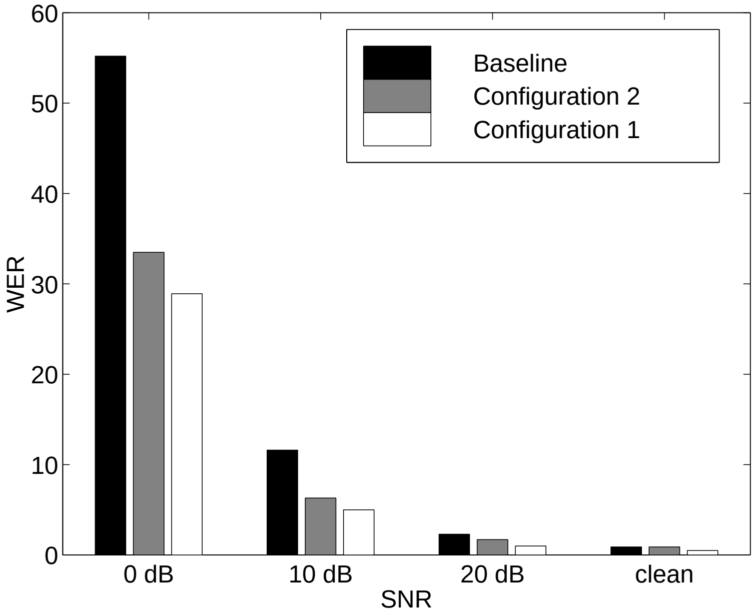

Fig. 13 – Test results (Aurora 2.0 connected digits data, test A = matched noise) for the HMM/ANN baseline (trained on clean speech) and two (a parameter-heavy (Configuration 1) and a parameter-light (Configuration 2)) narrow-band tandem (NB-tandem) HMM/ANN hybrids. All systems used J-RASTA-PLP features. Results from Table 1 in (Dupont and Ris, 2001).

Figure 13 (Dupont and Ris, 2001) shows that this multi-band tandem system employing robust J-RASTA-PLP features, with performance peaking at seven subbands, lead to a 50% relative error reduction over the baseline system under all noise conditions, including clean speech, even though its acoustic models were trained in noise.

5 Discussion

Multi-expert combination has deep roots in the theory of statistical estimation. We have certainly not even attempted to cover the broad range of techniques which could be said to belong to this field. The HMM/ANN hybrid models presented were the most successful models that were developed within

[18] The baseline system used for these tests is not the Aurora standard HMM/GMM, but an HMM/ANN. It was also trained on clean rather than multi-condition data. The results in Figure 13 are therefore not directly comparable with those in Figures 10 or 11.

the framework of the two European Community projects, SPHEAR[19] and RESPITE[20]. Closely related contemporary developments include (Cerisara, 1999a; Janin et al., 1999; Mirghafori, 1999; Dupont, 2000; Glotin, 2000; Jancovic and Ming, 2001; Shire, 2001). More distantly related work exploiting ANNs in ASR includes (Niles and Silverman, 1990; Bengio et al., 1992; Robinson, 1994). Other promising HMM/GMM based approaches to noise robust ASR include "missing data" (Green et al., 1995; Barker et al., 2001) and noise modelling approaches (Droppo et al., 2002) amongst many others.

It could be argued that all of the HMM/ANN systems reviewed have their conceptual roots in the early multi-band models (Bourlard and Dupont, 1996) which played the role of devil's advocate in making the clearly untenable assumption of (conditional) subband independence[21]. All of the models presented get around this independence assumption, but each model has its own strengths and weaknesses, which we summarise below.

5.1 Strengths and weaknesses of each system

All-combinations multi-band (AC MB)

While avoiding the assumption of subband independence, the AC MB model still assumes that each subband is either 100% reliable or 100% uninformative. As each data coefficient within a frequency subband can be independently clean or corrupted, the potential inaccuracy of this assumption increases with the width of each subband. Narrower subbands would be more accurate, but the number of subband combinations increases exponentially with the number of subbands..

All-combinations multi-stream (AC MS)

The AC MS model identifies each "subband" in the AC MB model with a separate representation of the fullband speech signal. In this approach any attempt to isolate noisy frequency subbands has been dropped, so all acoustic features are implicitly assumed to be clean, or to have matching noise conditions. The strength of this approach is that data from multiple "sensors" can be combined without assuming that the data from each sensor is independent (dependent streams are best processed together, but the number of free model parameters can be reduced, so improving performance, when genuinely independent streams are processed separately).

Noise-weighting of individual coefficients is currently not possible in ANN based systems. Reliability weighting of individual subbands is possible but not of great importance for AC MB or AC MS systems, because their performance has been found to depend much more on the method used for expert combination than the method used for weight estimation.

AC MS with hypothesis level combination

Static MAP weighting with the AC MS sum rule gave rise to hypothesis level MAP combination. This is a simple and principled rule which tests have shown can be effective for expert combination. However, insofar as it assumes that for each utterance the expert for just one stream combination is reliable and all others should be ignored, it does not permit the pooling of expert "opinions" which can also improve estimation performance.

Multi-stream tandem (MS-tandem)

The "tandem" approach, whereby ANNs are used for NLDA feature processing prior to input to a standard HMM/GMM or HMM/ANN system, also enables multi-stream combination while avoiding the assumption of stream independence. As with AC MS, the MS-tandem model is best suited for preprocessing streams which are already noise free, or have matching train/test noise conditions. This model was found to have world beating performance for the Aurora small vocabulary test under matched noise conditions, though not under mismatched conditions or with LVASR.

Narrow-band tandem (NB-tandem)

The NB-tandem model pre-processes narrow subbands so that any noise encountered during recognition will resemble the white noise used for training. On the other hand, if the bands are too narrow,

[19] http ://www.dcs.shef.ac.uk/~pdg/sphear/sphear.htm

[20] http ://www.dcs.shef.ac.uk/research/groups/spandh/projects/respite

[21] The assumption of class-conditional independence is far less inaccurate than would be an assumption of full independence.

speech will also resemble white noise and the posteriors "features" will carry no speech information. The test results presented here show that this system can give a strong advantage under mismatched noise conditions (none of the test noises resembled white noise), even when the original input features are relatively noise robust.

5.2 New directions

All of the hybrid systems presented are multi-stage systems whose performance will improve each time one of its component modules is improved. Due to the complexity of these systems, a large number of issues arise concerning their future development, the most important of which we discuss in this section.

Posteriors versus likelihoods estimation

The advocacy of ANNs as opposed to GMMs, although originally based on empirical evidence that ANNs using MLPs perform better than GMMs as classifiers[22], should be viewed more accurately as the advocacy of posteriors- over likelihoods based modelling, rather than of MLPs over GMMs. Minimum Bayes risk (i.e. minimum error) classification will always require posterior probability maximization and although the "class" to be recognised in ASR is the whole word sequence, so long as decoding is based on the initial recognition of sub-word speech units, and these are modelled by HMM states, then posterior probabilities are required for HMM states. While it is true that class posteriors can be obtained by applying Bayes rule to likelihoods modelled by GMMs trained with MCE, it is simpler to model class posteriors directly using ANNs, and by the principle of Occam's razor, simpler is better.

ANNs versus GMMs

One of the main criticisms of the use of ANNs for noise robust ASR is that uncertainty about the value of individual coefficients input to an ANN cannot be treated "in a theoretically sound probabilistic manner". However, classifier GMMs and MLPs have the same theoretical status as semi-parametric models for estimating class posterior probabilities. If the value of some of the acoustic features is uncertain (hence probabilistic), then Bayes optimum posteriors estimates are given by the expected value of the classifier outputs (Morris et al., 1998) whatever model is used. While it is true that this expectation integral is currently feasible for GMMs and not for MLPs, there is no reason why some other type of ANN could not be used for this purpose in future (Morris et al., 2000).

Noise removal from waveform

Microphone sampling frequency should be sufficient so as not to discard information which could be used for the separation of target speech from other interfering sounds (8 kHz, though given for telephone speech, is highly suboptimal). Classical noise removal techniques can be very effective and should not be overlooked (especially when the auditory system gives us obvious clues, such as the fact that ears tend to occur in pairs). Microphone arrays and Wiener filters can be very effective for noise estimation (McCowan and Sridharan, 2001).

Spectral noise removal versus noise modelling

Noise estimation techniques are also often applied after frequency analysis. For the purpose of noise modelling in ASR this estimated noise is modelled rather than removed. This has the advantage of potentially retaining any knowledge about the accuracy of the noise estimate, while this important knowledge is normally lost after noise removal. However, when secondary feature processing is in use (such as DCT, PCA, LDA, quantization, or posteriors estimation or NLDA by ANN), it will usually be the best option to remove the estimated noise while still in the spectral domain, before it is spread over all of the secondary features. In this case it has been found that it is a good rule not to correct observations which you are not very sure need correcting ("partial imputation"), and for values which are to be corrected, various methods exist which give better results than simple spectral subtraction.

Multiple signal representations

[22] GMM fans may counter that the real reason ANNs were explored was more to do with wishful thinking about the mystical capabilities of "brain-like neural networks", that MLPs are not brain-like at all, and GMMs have a much better foundation in statistical modelling theory. As usual with many disputes, the truth probably lies somewhere in between.

Recognition can always be enhanced by combining multiple complementary signal representations. Such representations may arise from multiple sources of biometric data (typically facial image and speech signal for ASR, though on-line signature, iris, fingerprint and perhaps olfactory data would also be of use in speaker identification), as well as feature analysis at different spatial, temporal or frequency scales. Stream combination may be at the feature, state posteriors and/or utterance level.

Multi-condition training

Training with multi-condition data can greatly improve performance under matched conditions, but the wider the range of conditions used in training the flatter the distribution of data in each phonetic class, which can lead to lower performance under any one condition. The NB-tandem system (Section 4.2), which uses training in white noise, was the only system to significantly improve performance in clean speech (at least in the one test reported). The adverse flattening effect of multi-condition training may be offset to some extent by a reduction in over-fitting to uninformative detail, but the particular success of the NB-tandem approach may also be explained as follows. Log compressed data values at the spectral peaks for each HMM state class are far less affected by noise than values in spectral valleys. Training with multiple levels of white noise systematically "floods" all spectral valleys with high variance data, while peaks retain low variance. This will directly lead to flat "don't care" within-class probability density functions (pdfs) for all in-valley (ergo noise prone) data coefficients, whereas real noises, having non flat spectra, will tend to also flatten spectral peaks.

The systematic identification of "don't care" coefficients is an interesting topic for future attention, because the possibility of such coefficients is generally overlooked in multivariate pdf estimation, even though many scenarios come to mind in which class membership is best expressed as a union of incomplete, rather than complete, conjunctions of feature attributes (e.g. class $1 : x_1 = a$ and $x_2 = b$; class $2 : x_2 = c$; class $3 : x_1 = d$ or $x_2 = e$).

Variety of classifier architectures

With multi-expert systems, although individual expert performance is preferably accurate, it is more important that each classifier is unbiased and has complementary error characteristics. Every classifier reported here was an MLP. A classifier which often outperforms the MLP (especially with limited training data) is the support vector machine[23] (SVM), which was only recently developed for use with high dimensional data (Collobert and Bengio, 2001). However, high performance classifiers tend to have similar error characteristics. From this point of view the Gaussian RBF ("Radial Basis Function") ANN classifier (Bishop, 1995) may be a more interesting candidate, precisely because it tends not to perform as well as the MLP (Morris et al., 2000). Another suitably strange candidate may be a classifier based on the "product mixture of Gaussians" model (Hinton and Brown, 2001) (if this can get over its complexity problems).

New combination rules and weighting schemes

Multi-stream system performance is much more sensitive to changes in system architecture and/or combination rules than to different weighting strategies. It is conceivable that new schemes for posteriors combination could be leveraged from the idea which lies behind the "probabilistic union model", which up until now has only been applied in a somewhat ad hoc way to subband likelihoods combination (Jancovic and Ming, 2001). The AC sum rule requires that the indicator events $b^{(i)}$ in (6) are both mutually exclusive and exhaustive, while the union model (being based on an inclusive rather than an exclusive OR of indicator events) would only require that they were exhaustive.

Asynchronous decoding

When combining evidence from data streams which are not frame synchronous, rather than render them synchronous by interpolation, performance can sometimes be gained by allowing asynchrony (Cerisara, 1999a; Mirghafori and Morgan, 1999; Cerisara et al., 2000; Bengio, 2003). This increases the size of the search space, but not necessarily beyond practical limits, providing the time lag between streams is limited.

One-stage multi-expert training

[23] SVM outputs do not have a direct interpretation as probabilities and have to be transformed so that they are all positive and sum to one, but this can be easily arranged.

All of the multi-expert systems reported are multi stage processes involving at least three processing stages. Discriminative feature analysis, state probabilities (likelihoods or posteriors) modelling and weighted expert combination all require training. While training is normally done on a modular basis, efforts have been made to improve performance by one-stage training (Cerisara, 1999b; Sharma, 1999). So far this has not lead to significant performance improvements compared to the improvements achieved by the systems we have reported which use modular training. However, some models are more naturally suited than others to one-stage training. It may be worth experimenting with "gated mixture of experts" models, in either their simple (Jacobs et al., 1991) or hierarchical (Jordan and Jacobs, 1994) form. These are interesting brain-like models (synaptic gating occurs extensively throughout the brain) which could be used as one stage classifiers. They have been extensively developed theoretically but not yet been tested in speech recognition.

Robust ASR is a complex problem which cannot be solved in one step. Each of the separate steps of noise estimation, through discriminative feature analysis, model adaptation, posteriors or likelihoods estimation, and expert combination, to decoding, must select the model which is best suited both to this processing stage and to the specific ASR application. The models presented make a number of new additions to our toolbox, most notably in the areas of robustness to mismatched noise, and multi-modal feature combination.

6 Conclusion

We have reviewed a number of models for multi-modal data fusion in which combination takes place at one or more of the levels of input features, state posteriors or utterance hypothesis. The theoretical advantages and limitations of each system were discussed.

All of the results presented were on connected digits recognition under clean or matched noise conditions. Most tests compare against an HMM/ANN baseline whose performance compares to a state-of-the-art HMM/GMM. Test results comparing these systems directly against each other or on a global scale were regrettably lacking at the time of publication.

In the one test available where an MS HMM/ANN system's performance is compared on a global scale (connected digits recognition under matched noise conditions, Figure 11) the MS-tandem system comes out on top. While this result was later found not to extend to the mismatched noise case, the NB-tandem appears to overcome the problem of noise mismatch.

Most of the models reviewed were HMM/ANN based, but as likelihoods can always be converted to posteriors using Bayes' rule, all of the equations on which these models are based can be exploited equally by likelihoods based (i.e. HMM/GMM) and posteriors based (i.e. HMM/ANN) systems. We are now in the position that MLPs are better suited than GMMs to posteriors estimation, while GMMs are better suited than MLPs to adaptive noise (and speaker) modelling. This means that until the day when ANNs are developed that can compete with GMMs for ease of model adaptation, we must compromise between systems using (i) inferior feature level noise compensation, followed by superior posteriors modelling, and (ii) superior noise model adaptation with inferior likelihood based models.

We hope that by collecting together and presenting this range of multi-expert HMM/ANN models, some of which are very recent and will not be familiar to many readers, we will encourage others to further test their effectiveness in a wider range of noise conditions, and large vocabulary recognition tasks, and to overcome some of the limitations which we have identified.

A Proof that MAP sum weights select maximum posterior

The weighted sum $P(Q|X) = \sum_i w_i P(Q|X^{(i)})$ has the form $A = \sum_{i=1}^{i=M} w_i a_i$, where a_i are fixed values. We can find w to maximise this, subject to the constraints $\sum_i w_i = 1$ and $w_i \geq 0$, as follows. First, without loss of generality, label a_i (which are all positive) in order of decreasing magnitude.

$$A = w_1 a_{\text{max}} + w_2 a_2 + \ldots + (1 - w_1 - \ldots) a_{\text{min}} \tag{17}$$

Differentiating with respect to each free parameter w_j $(j = 1, \ldots, (M-1))$, gives

$$\frac{dA}{dw_j} = a_j - a_{\text{min}} \tag{18}$$

But $a_j - a_{\text{min}} \geq 0$, so A is always increasing with each w_j, and increases fastest with increase in w_1. From this it follows that A is maximised when $w_1 = 1$ and all other $w_i = 0$. Therefore

$$\max_w A = \max_w \sum_i w_i a_i = a_{\text{max}} = \max_i P(Q|X^{(i)}) \tag{19}$$

Acknowledgments

The authors would like to thank Hérvé Bourlard, Stéphane Dupont, Dan Ellis and Christophe Ris for providing the experimental details and results for their HMM/ANN hybrid systems which were amongst those described in this article. This work was supported by the Swiss Federal Office for Education and Science (OFES) in the framework of both the EC/OFES SPHEAR (SPeech, HEAring and Recognition) project and the EC/OFES RESPITE (REcognition of Speech by Partial Information TEchniques) project. Additionally, this work was sponsored by the Portuguese FCT (Fundação para a Ciencia e a Tecnologia) scholarship SFRH/BPD/6757/2001.

Références

Allen, J., 1994. How do humans process and recognise speech ? IEEE Transactions on Speech and Audio Processing 2 (4), 567–577.

Barker, J., Green, P., Cooke, M., 2001. Linking auditory scene analysis and robust ASR by missing data techniques. In : Proc. Workshop on Innovation in Speech Processing (WISP) 2001. Stratford-upon-Avon, UK, pp. 295–307.

Barker, J., Josifovski, L., Cooke, M., Green., P., 2000. Soft decisions in missing data techniques for robust automatic speech recognition. In : ICSLP 2000. pp. 373–376.

Bengio, S., 2003. An asynchronous hidden markov model for audio-visual speech recognition. In : Becker, S., Thrun, S., Obermayer, K. (Eds.), Advances in Neural Information Processing Systems, NIPS 15. MIT Press, pp. 1237–1244.

Bengio, Y., Mori, R. D., Flammia, G., Kompe, R., 1992. Global optimisation of a Neural Network - Hidden Markov Model hybrid. IEEE Trans. on Neural Networks 3 (2), 252–259.

Bishop, C., 1995. Neural Networks for Pattern Recognition. Clarendon Press, Oxford.

Bourlard, H., Dupont, S., 1996. A new ASR approach based on independent processing and recombination of partial frequency bands. In : ICSLP '96. pp. 426–429.

Bourlard, H., Dupont, S., 1997. Subband-based speech recognition. IEEE Trans. on Acoustics, 1251–1254.

Bourlard, H., Dupont, S., Ris, C., 1996. Multi-stream speech recognition. Technical report IDIAP RR-07, IDIAP, Martigny, Switzerland.

Bourlard, H., Morgan, N., 1994. Connectionist Speech Recognition – A Hybrid Approach. Kluwer Academic Publishers, 101 Philip Drive, Assinippi Park, Norwell, Massachusetts 02061 USA.

Cerisara, C., 1999a. Computational Models of Speech Pattern Processing. Springer Verlag in cooperation with NATO ASI Series, Berlin, Ch. Dealing with loss of synchronism in multi-band continuous speech recognition systems, pp. 90–95.

Cerisara, C., 1999b. Contribution de l'approche multi-bande a la reconnaissance automatique de la parole. Ph.D. thesis, Institut National Polytechnique de Lorraine, Nancy, France.

Cerisara, C., Fohr, D., Haton, J.-P., 2000. Asynchrony in multi-band speech recognition. In : ICASSP 2000. pp. 1121–1124.

Cerisara, C., Haton, J.-P., Mari, J.-F., Fohr, D., 1998. A recombination model for multi-band speech recognition. In : ICASSP '98. pp. 717–720.

Cole, R., Noel, M., Lander, T., Durham, T., 1995. New telephone speech corpora at CSLU. In : Eurospeech '95. pp. 821–824.

Collobert, R., Bengio, S., 2001. Svmtorch : support vector machines for large-scale regression problems. Machine Learning Research 1, 143–160.

Davis, S. B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoustics, Speech, and Signal Processing ASSP-28 (4), 357–366.

Droppo, J., Acero, A., Deng, L., 2002. Uncertainty decoding with splice for noise robust speech recognition. In : ICASSP '02. pp. 57–60.

Duda, R., Hart, P., 1973. Pattern Classification and Scene Analysis. John Wiley and Sons.

Dupont, S., 2000. Etudes et developpement de nouveaux paradigmes pour la reconnaissance robuste de la parole. Ph.D. thesis, Laboratoire TCTS, Universite de Mons, Mons, Belgium.

Dupont, S., Luettin, J., 1998. Using the multi-stream approach for continuous audio-visual speech recognition : experiments on the M2VTS database. In : ICSLP '98. pp. 1283–1286.

Dupont, S., Ris, C., 2001. Multi-band with contaminated training data. In : Eurospeech '01. pp. 429–432.

Ellis, D., October 2002. Personal communication.

Ellis, D., Reyes-Gomez, M., 2001. Investigations into Tandem acoustic modeling for the Aurora task. In : Eurospeech '01. pp. 189–192.

Evermann, G., Woodland, P., 2000. Posterior probability decoding, confidence estimation and system combination. In : Speech Transcription Workshop.

Fiscus, J., 1997. A post-processing system to yield reduced word error rates : Recogniser output voting error reduction (ROVER). In : Proc. IEEE ASRU Workshop, Santa Barbara. pp. 347–352.

Fletcher, H., 1953. Speech and Hearing in Communication. Krieger, New York.

Fontaine, V., Ris, C., Boite, J.-M., 1997. Nonlinear discriminant analysis for improved speech recognition. In : Eurospeech '97. pp. 2071–2074.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition, 2nd Edition. Academic Press, Boston.

Ghitza, O., 1994. Auditory models and human performance in tasks related to speech coding and speech recognition. IEEE Transactions on Speech and Audio Processing 2 (1), 115–131.

Glotin, H., 2000. Élaboration et comparaison de systèmes adaptatifs multi-flux de reconnaissance robuste de la parole : incorporation des indices d'harmonicité et de localisation. Ph.D. thesis, Institut National Polytechnique de Grenoble, Grenoble, France.

Glotin, H., Berthommier, F., 2000. Test of several external posterior weighting functions fur multiband full combination ASR. In : ICSLP 2000. pp. 333–336.

Goldberg, R., Riek, L., 2000. A Practical Handbook of Speech Coders. CRC Press, Boca Raton, Florida.

Grant, K., Braida, L., 1991. Evaluating the articulation index for auditory-visual input. JASA 89 (6), 2952–2960.

Green, P. D., Cooke, M. P., Crawford, M. D., 1995. Auditory scene analysis and hidden Markov model recognition of speech in noise. TransSP, 401–404.

Greenberg, S., Kingsbury, B. E. D., 1997. The modulation spectrogram : In pursuit of an invariant representation of speech. In : ICASSP '97. pp. 1647–1650.

Haeb-Umbach, R., Ney, H., 1992. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In : ICASSP '92. pp. 13–16.

Hagen, A., 2001. Robust speech recognition based on multi-stream processing. Ph.D. thesis, Département d'Informatique, École Polytechnique Fédérale de Lausanne, Switzerland.

Hagen, A., Morris, A., Bourlard, H., 1999. Different weighting schemes in the full combination subbands approach in noise robust ASR. In : Proc. ESCA Workshop on Robust Methods for Speech Recognition in Adverse Conditions. pp. 199–202.

Hagen, A., Morris, A., Bourlard, H., 2000. From multi-band full combination to multi-stream full combination processing in robust ASR. In : ISCA ITRW Workshop on Automatic Speech Recognition – Challenges for the new millenium (ASRU2000). pp. 175–180.

Hagen, A., Neto, J., 2003. Multi-stream processing using context-independent and context-dependent hybrid systems. In : ICASSP '03. pp. 277–280.

Halberstadt, A., Glass, J., 1998. Heterogeneous measurements and multiple classifiers for speech recognition. In : ICSLP '98. pp. 995–998.

Hansen, L., Salamon, L., 1990. Neural network ensembles. IEEE Transactions of Pattern Analysis and Machine Intelligence 12 (10), 993–1001.

Heckmann, M., Berthommier, F., Kroschel, K., 2001. Optimal weighting of posteriors for audio-visual speech recognition. In : ICASSP '01. pp. 161–164.

Hermansky, H., April 1990. Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustical Society of America 87 (4), 1738–1752.

Hermansky, H., Ellis, D., Sharma, S., 2000. Tandem connectionist feature extraction for conventional HMM systems. In : ICASSP 2000. pp. 1635–1638.

Hermansky, H., Morgan, N., Bayya, A., Kohn, P., 1992. RASTA–PLP speech analysis technique 1, 121–124.

Hermansky, H., Tibrewala, S., Pavel, M., 1996. Towards ASR on partially corrupted speech. In : ICSLP '96. pp. 462–465.

Hinton, G., Brown, A. D., 2001. Training many small hidden markov models. In : Proc. Institute of Acoustics workshop WISP 2001, Workshop on Innovation in Speech Processing. pp. 1–17.

Hirsch, H.-G., Pearce, D., 2000. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In : ISCA ITRW Workshop on Automatic Speech Recognition – Challenges for the new millenium (ASRU2000). pp. 181–188.

Hochberg, M., Cook, G., Renals, S., Robinson, A., Schechtman, R., 1995. ABBOT hybrid connectionist-hmm large-vocabulary recognition system. In : Spoken Language Systems Technology Workshop. pp. 170–176.

Houtgast, T., Steeneken, H., 1985. A review of the mtf concept in room acoustics and its use for speech intelligibility. JASA 77, 1069–1077.

Houtgast, T., Verhave, J., 1991. A physical approach to speech quality assessment : Correlation patterns in the speech spectrogram. In : Eurospeech '91. pp. 285–288.

Jacobs, R. A., Jordan, M. I., Nowland, S. J., Hinton, G. E., 1991. Adaptive mixtures of local experts. Neural Computation 3, 78–87.

Jancovic, P., Ming, J., 2001. A multi-band approach based on the probabilistic union model and frequency-filtering features for robust speech recognition. In : Eurospeech '01. pp. 579–582.

Janin, A., Ellis, D., Morgan, N., 1999. Multi-stream speech recognition : Ready for prime time ? In : Eurospeech '99. pp. 591–594.

Jordan, M. I., Jacobs, R. A., 1994. Hierarchical mixture of experts and the EM algorithm. Neural Computation 6, 181–214.

Kirchhoff, K., 1998. Combining articulatory and acoustic information for speech recognition in noisy and reverberation environments. In : ICSLP '98. pp. 891–894.

Kirchhoff, K., Fink, G., Sagerer, G., 2000. Conversational speech recognition using acoustic and articulatory input. In : ICASSP 2000. pp. 1435–1438.

Lippmann, R. P., 1996. Accurate consonant perception without mid-frequency speech energy. IEEE Transactions on Speech and Audio Processing 4 (1), 66–69.

McCowan, I., Sridharan, S., 2001. Microphone array sub-band speech recognition. In : ICASSP '01. pp. 185–188.

Mirghafori, N., 1999. A multi-band approach to automatic speech recognition. Ph.D. thesis, ICSI, Berkely, California.

Mirghafori, N., Morgan, N., 1998. Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers. In : ICSLP '98. pp. 743–746.

Mirghafori, N., Morgan, N., 1999. Sooner or later : Exploring asynchrony in multi-band speech recognition. In : Eurospeech '99. pp. 595–598.

Morris, A., Cooke, M., Green, P., 1998. Some solutions to the missing features problem in data classification, with application to noise robust ASR. In : ICASSP '98. pp. 737–740.

Morris, A., Hagen, A., Bourlard, H., 2001a. MAP combination of multi-stream HMM or HMM/ANN experts. In : Eurospeech '01. pp. 225–228.

Morris, A., Hagen, A., Glotin, H., Bourlard, H., 2001b. Multi-stream adaptive evidence combination to noise robust ASR. Speech Communication 34 (1-2), 25–40.

Morris, A., Josifovski, L., Bourlard, H., Cooke, M., Green, P., 2000. A neural network for classification with incomplete data : Application to robust ASR. In : ICSLP 2000. pp. 345–348.

Morris, A., Payne, S., Borlard, H., 2002. Low cost duration modelling for noise robust speech recognition. In : ICSLP '02. pp. 1025–1028.

Niles, L., Silverman, H., 1990. Combining Hidden Markov Models and Neural Network classifiers. In : ICASSP '90. pp. 417–420.

Ripley, B. D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge, UK.

Robinson, T., 1994. The application of recurrent Neural Nets to phone probability estimation. IEEE Transactions on Neural Networks, 829–832.

Rogozan, A., Deléglise, P., 1998. Adaptive fusion of acoustic and visual sources for automatic speech recognition. Speech Communication 26 (1–2), 149–161.

Rumelhart, D., Hinton, G., Williams, R., 1986. Learning internal representations by error propogation. In : Rumelhart, D., McClelland, J. (Eds.), Parallel Distributed Processing. Exploration of the Microstructure of Cognition. Vol. 1 : Foundations. MIT Press, pp. 318–362.

Sharma, S., October 1999. Multi-stream approach to robust speech recognition. Ph.D. thesis, Oregon Graduate Institute of Science and Technology, Oregon, USA.

Shire, M. L., 2000. Discriminant training of front-end and acoustic modeling stages to heterogeneous acoustic envrionments for multi-stream automatic speech recognition. Ph.D. thesis, University of California, Berkeley, USA.

Shire, M. L., 2001. Multi-stream ASR trained with heterogeneous reverberant environments. In : ICASSP '01. pp. 253–256.

Silipo, R., Greenberg, S., Arai, T., 1999. Speech intelligibility derived from exceedingly sparse spectral information. In : Eurospeech '99. pp. 2687–2690.

Steeneken, H., Houtgast, T., 1980. A physical method for measuring speech transmission quality. JASA 67 (1), 318–326.

Steeneken, H., Houtgast, T., 1999. Mutual dependence of the ocatve-band weights in predicting speech intelligibility. Speech Communication 28, 109–123.

Tibrewala, S., Hermansky, H., 1997. Sub-band based recognition of noisy speech. In : ICASSP '97. pp. 1255–1258.

Tomlinson, M. J., Russel, M. J., Brooke, N., 1996. Integrating audio and visual information to provide highly robust speech recognition. In : ICASSP '96. pp. 821–824.

Tumer, K., Ghosh, J., 1996. Analysis of decision boundaries in linearly combined neural classifiers. Pattern Recognition 29 (2), 341–348.

Varga, A., Steeneken, H., Tomlinson, M., Jones, D., 1992. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical report, DRA Speech Research Unit, Malvern, England.

Warren, R., Riener, K., Bashford, J., Brubaker, B., 1995. Spectral redundancy : Intelligibility of sentences heard through narrow spectral slits. Perception and Psychophysics 57 (2), 175–182.

Wu, S., Kingsbury, B., Morgan, N., Greenberg, S., 1998. Incorporating information from syllable-length time scales into automatic speech recognition. In : ICASSP '98. pp. 721–724.