Concluding Remarks on Multi-band and Multi-stream Research for Noise-robust ASR

Astrid Hagen[1,2]

[1] INESC-ID, Rua Alves Redol 9, Lisbon, Portugal
astrid.hagen@l2f.inesc-id.pt
[2] formerly at: IDIAP, Rue du Simplon 4, Martigny, Switzerland

Abstract. This paper summarizes our work carried out over the last four years on multi-band and multi-stream processing as means to achieve robust automatic speech recognition (ASR). Main focused is laid on the "full combination" approach which integrates over all possible positions of reliable data, instead of combining the nominal subbands or streams only. The main combination rules are summarized and the final results from a wide range of experiments are presented.

1 Introduction

Different approaches to increase the robustness of an automatic speech recognizer comprise, among others, more appropriate feature extraction, better acoustic modeling and advanced decoding schemes. In this framework the goal of this work was to investigate and develop new paradigms for noise robust ASR based on multi-band (MB) and multi-stream (MS) processing[3].

After the speech signal has been converted to the spectral domain, in **MB** processing, the entire frequency domain is split into frequency subbands which are processed independently up to a certain point where the information from each band is recombined. In **MS** processing, either the entire frequency domain is considered several times, each time employing different processing strategies, or other modalities, such as visual representations, of speech production are included. The information from each of these streams is correspondingly recombined later in the process.

Both approaches try to better utilize the inherent redundancy in the speech signal either by processing different parts of the signal separately or by different processing of the same signal stream. If the streams are correlated, it can be assumed that combination is best carried out on the feature level so that dependencies between the streams can be modeled. In case when the streams are corrupted by noise, the correlation between the streams is decreased. It can thus be assumed that the streams are better modeled independently, as this is likely to result in independent errors conducted by each stream recognizer due

[3] Although the latter is a generic term of the first, we distinguish these two approaches due to historical reasons: MB processing was – as far as our research is concerned – investigated first and its principle was then generalized to multi-stream processing.

to train/test mismatch. Nothing can be done about these errors when dealing with a single-stream (fullband) recognizer only. However, when combining the outputs of two or more recognizers, independent errors coming from any one of them can be dampened. Thus, the MB and MS systems are expected to provide higher noise robustness to any kind of noise than a single-stream system, without any knowledge of the noise or the necessity of different training databases and noise adaptation phase.

In this framework, we investigated several frame-level **combination approaches**, some of which employ a reliability term for each subband or stream [9]. The MB and MS strategies were developed on clean speech data and their noise robustness was tested and evaluated on noise-corrupted speech with the noise stemming from various additive noise environments. The different MB and MS recognizers were compared amongst each other as well as to the baseline fullband recognizers.

Our research was carried out in the framework of Hidden Markov Model (HMM) based speech recognizers, where HMM emission probabilities were estimated through either Gaussian Mixture Models (GMMs), or Artificial Neural Networks (ANNs). The former outputs likelihoods so that combination of different streams is carried out on these, whereas the latter outputs posteriors which are used for recombination in this case (here on the frame level). After recombination of posteriors, the recombined posteriors are divided by the prior probabilities to obtain (scaled) likelihoods for the regular ("one-dimensional") Viterbi decoder as used here.

2 Multi-Band Processing

In *multi-band* processing, the speech signal in the spectral domain is split into several subbands which are processed independently for feature extraction and possibly probability estimation before they are recombined for further processing. In the case when noise only occurs in one frequency subband, it does therefore not mix with the other clean feature coefficients which allow for reliable decoding of the clean part of the speech. Similarly, in *missing data* (MD) [4, 3, 12] processing as applied to robust ASR, it is tried to segregate speech and noise in the input signal, and then to recognize at each time frame the clean speech part only. This includes the necessity for a noise detection algorithm and for the processing of continuously varying combinations of (clean) feature coefficients. Moreover, only one fixed decomposition into clean and noisy data (a so-called "MD mask") is considered at each time frame.

Original subband processing misses important frequency correlation information among subbands and is therefore usually not competitive in real-environmental noise. We developed in this work, an approach to subband processing which provides a solution to the problem of both loss of frequency correlation in MB processing and fixed MD masks through a revised decomposition of the frequency band into an *exhaustive and mutual exclusive set of frequency subbands*. This induces new combination strategies as described below.

Full Combination Processing In MB ASR it was up to now assumed that subbands could be processed independently, with each subband modeled by a distinct recognizer. In the case of noise-corrupted speech in one subband, correct recognition on the remaining clean subbands could then provide enough information to decode the entire input data. In case of clean speech and speech corrupted with wideband noise, however, experiments in ASR have shown that a MB system of this type very often leads to decreased performance as compared to a fullband recognizer, due to missing cross correlation information. To model more closely what is actually going on in humans who integrate information from even dispersed frequency regions, and to obtain higher performance in both clean and (wideband) noise corrupted speech by a MB system, we had to find a revised model which also exploits correlation information between (adjacent and non-adjacent) subbands. This should be done by integrating also dispersed frequency information, when some frequency regions are missing, in order to exploit this correlation and redundancy in the spectrum.

Thus, at each time frame, as much clean correlated information as possible should be modeled. In the MD approach, noise corrupted frequencies in each frame are detected and excluded, while the remaining reliable data is modeled as a single stream. However, accurate noise detection is very difficult. In the "full combination" (FC) approach taken here, data is divided into subbands and recognition is performed on *every possible combination of subbands*, after which the output from these experts is integrated by one of several possible combination strategies.

The FC paradigm for MB ASR For most application areas, the position of the noise is not known and can be in any subband and any number of subbands. We therefore have to find a way in which we can consider all possible subsets of the frequency domain in order to find the clean data set.

For this, let us define the set of all possible combinations of B subbands, which include the streams consisting of no, one, two etc. (adjacent and non-adjacent) subbands up to the combination of all subbands, as C, and the set of events b_i $(i = 1, \ldots, \mathcal{B} = 2^B)$ as follows:

> *$\mathcal{B}$ denotes the set of events b_i that data in combination i is clean speech data, and data not in combination i is completely uninformative and can therefore be regarded as missing.*

On the assumption that each subband is either completely clean or completely uninformative, such a set of events is mutually exclusive and exhaustive, as only one combination of subbands can be the largest clean combination, and one or other must be the true clean combination, because all possible combinations have been considered. Denoting $P(b_i)$ the probability that event b_i occurs, we can write:

$$\begin{aligned} P(\cup_i b_i) &= \sum_{i=1}^{\mathcal{B}} P(b_i) \qquad \text{(mutually exclusive)} \\ &= 1 \qquad\qquad\qquad\qquad\quad \text{(exhaustive)} \end{aligned} \tag{1}$$

If some subbands are not corrupted by noise, it is likely that the best stream[4] is the largest combination of clean subbands[5]. However, under wideband noise conditions it can also be the case that some less noisy subset carries more useful information than the empty set.

Let us now consider how this new FC approach to subband processing can be implemented in a speech recognizer. Considering all possible combinations of subbands means that features have to be extracted not only in the nominal subbands but also in each combination of subbands, i.e. in the $\mathcal{B}$ feature streams (note that this includes the empty set). Data within each feature stream can be further processed for decorrelation and/or other transformations, as required.We can then associate with each event b_i an expert i which has at its input the clean data defined by event b_i. In (posterior-based) probability combination, a recognizer has thus to be trained on each of the $\mathcal{B}$ feature sreams, as shown in **Figure 1** for the case of two subbands. Realization of the FC approach in posterior- and likelihood-based systems is discussed below. A further advantage

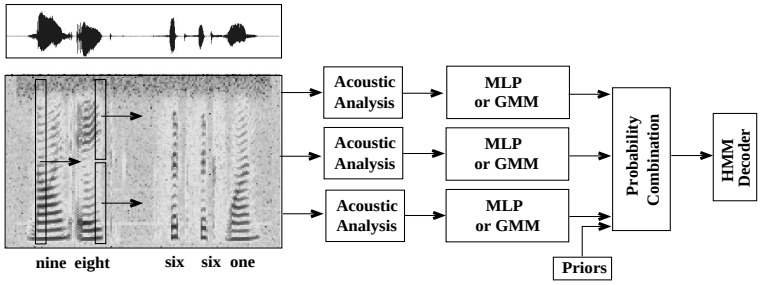

Fig. 1. Illustration of full combination processing with MLP or GMM classifiers for two subbands. Features are extracted from all possible combinations of subbands.

of FC processing over "standard" subband processing is that the question of how many subbands are to be chosen and the exact position of the subbands gets less important as in the FC approach all subbands are considered by themselves *and in combination* and thus correlation between all subbands is considered.

3 Combination Strategies

Different strategies for the recombination of stream probabilities exist, such as the sum or product rule. In each of them, we have to combine likelihoods or

[4] Note that, in order to be able to refer to a 'subband ' and 'combination of subbands' in one term, we use the term '(*data*) stream' to account for both.

[5] This is under the assumption that the stream acoustic models are trained on clean speech only.

posteriors according to a function which often depends on weights representing the reliability of each (subband) stream.

In this section, we present the probability combination strategies which were developed in the framework of this work and which are based on the FC approach introduced in the preceding section. Each of the combination strategies is presented for both the posterior-based case and the likelihood-based case. The likelihoods need to be normalized before combination to account for the different range they usually cover.

Despite the fact that the preceding section was concerned with combining subband experts, the expert combination strategies presented here are not specific to subband expert combination, but can be applied to combinations of experts trained on any (preferably complementary) data streams. More specifically, the combination strategies are also an important part of the MS approach.

3.1 FC sum rule

FC posterior decomposition For posterior decomposition, a separate expert is trained for each of the $\mathcal{B}$ possible combinations. Introducing the hidden variable b_i $(i = 1, \ldots, \mathcal{B})$ indicating which band subset is clean, as defined in Section 2, and with the b_i's being mutually exclusive and exhaustive, $P(q_k|x)$ can be expressed as

$$P(q_k|x) = \sum_{i=1}^{\mathcal{B}} P(q_k|x_i)P(b_i|x) \quad \text{by definition of } b_i \tag{2}$$

$P(b_i|x)$ is the *reliability term* for each expert. If b_i is true, then $P(q_k|x_i)$ should be accurately estimated by expert i (which was trained on clean data). Otherwise the estimate will not be reliable.

Approximation to Full Combination (AFC) In the case of posterior-based experts (such as Multi-Layer Perceptrons (MLP)), it is necessary to train 2^B (MLP) experts, and the approach is thus limited to a small number of subbands. We, therefore, proposed an approximation scheme which estimates the probabilities for each *combination* of bands based on the *single* band experts only.

Under the assumption of conditional independence between subbands l in a combination x_i $p(x_i|q_k) \simeq \prod_{l \in x_i} p(x_l|q_k)$, we can derive the posteriors $P(q_k|x_i)$ for each subband-combination from the single-subband posteriors $P(q_k|x_l)$ in this combination (i.e. $l \in x_i$) as follows

$$P(q_k|x_i) = \Theta\, P^{1-|x_i|}(q_k) \prod_{l \in x_i} P(q_k|x_l) \tag{3}$$

where Θ is a normalization constant independent of q_k, such that $\sum_{k=1}^{K} P(q_k|x) = 1$ [8]. These approximated combination posterior probabilities (3) can now be used in any combination strategy where separately trained posteriors are used, such as (2). We see in the following how with FC for likelihoods, under certain conditions the stream likelihoods can be derived from the fullband likelihood without training other than the fullband expert.

FC likelihood decomposition using marginalization We can convert the FC SUM rule for posteriors (2) to a FC SUM rule for likelihoods by using Bayes' rule.

$$\frac{p(x|q_k)}{p(x)} = \sum_{i=1}^{\mathcal{B}} \frac{p(x_i|q_k)}{p(x_i)} P(b_i|x) \tag{4}$$

where $p(x_i) = \sum_{k=1}^{K} p(x_i|q_k)P(q_k)$.

In the FC SUM rule for likelihoods (4), we sum over all possible positions ($i = 1, \ldots, \mathcal{B}$) of reliable subbands. Under the condition that subband combination coefficients are selected from fullband coefficients without further processing (such as orthogonalization within a combination), the parameters for the marginal probability density functions (pdfs) $p(x_i|q_k)$ can be obtained directly from the parameters for the fullband pdf by marginalization.

Following the derivation which leads to expression of the marginal pdf for the data "present" in MD processing [4, 9], we can derive the state likelihoods $p(x_i|q_k)$ for each stream i by integrating over the unreliable, that is, "missing" part $x'_i = x - x_i$ of the data, which is disregarded in the respective stream:

$$p(x_i|q_k) = \int_{x'_i} p(x|q_k)\, dx'_i \tag{5}$$

For the mixture pdfs of M mixtures m_j as commonly used for likelihood modeling it holds:

$$\int_{x'_i} p(x|q_k)\, dx'_i = \sum_{j=1}^{M} P(m_j|q_k) \prod_{l \in s_i} p(x_l|m_j, q_k) \tag{6}$$

where s_i denotes the set of feature coefficients in subband combination i. In the case where each mixture component pdf $p(x|m_j, q_k)$ is modeled as a diagonal covariance Gaussian, with mean μ_{jk} and variance vector σ^2_{jk}, the mean and variance vectors for the marginal pdf $p(x_l|m_j, q_k)$, i.e. μ_{ljk} and σ^2_{ljk}, are simply obtained by striking out the rows and columns from the mean vector μ_{jk} and covariance matrix σ^2_{jk} corresponding to the missing components [4]. Substituting (6) back into (4) we get the full combination formula using marginalization for likelihood-based systems.

In the case when each stream only comprises one feature component, the above implementation of the FC approach can be interpreted in MD terminology as *a weighted sum over all possible sets of hard MD masks* using marginalization without bounds.

Preliminary experiments employing marginalization in FC MB ASR revealed that although this avoids the need to train more than one fullband expert, the remaining problem of having to evaluate the marginal likelihood for every combination of subbands is still very computationally expensive, and this prevented us from running further experiments.

3.2 FC product rule

Experimental results have often shown that, despite the limitations of the inaccurate independence assumption between the different recognizers working on each combination of subbands, the recombination by a product can be a more effective method of combining the outputs of multiple classifiers than the sum rule [6, 1, 5, 7, 10, 11].

FC product rules for likelihoods Under the inaccurate assumption of independence between the different recognizers, the full likelihood can be decomposed into a product of $\mathcal{B}$ stream likelihoods for each state q_k, according to:

$$p(x|q_k) \simeq \Theta_k \prod_{i=1}^{\mathcal{B}} p^{w_i}(x_i|q_k) \tag{7}$$

with $p(x_i|q_k)$ the state likelihood of expert i, which was trained on part x_i of data x only, and $\Theta_k = \frac{1}{\prod_i \theta_{ik}}$ a normalization constant, where $\theta_{ik} = \int_{x_i} p^{w_i}(x_i|q_k)\, dx_i$ so that $\int p(x|q_k)\, dx = 1$.

FC product rules for posteriors Under the assumption of conditional independence used in (7), we can derive for the posterior-based case the FC PRODUCT rule as follows:

$$P(q_k|x) = \Theta\, P^{1-\mathcal{B}}(q_k) \prod_{i=1}^{\mathcal{B}} P(q_k|x_i) \tag{8}$$

where Θ is a normalization constant, independent of q_k, such that $\sum_k P(q_k|x) = 1$ (for exact derivation see [9, p. 87]).

4 Multi-Stream Processing

In *multi-stream* processing, different possibilities exist to incorporate additional knowledge sources. They can stem, among others, from different data recordings (such as audio and visual streams), pre-processing, feature extraction, or from a different choice, structure and training of the classifiers. In this work, we concentrated on the use of different *feature* streams, from either different feature extraction techniques (such as PLP and MFCC features) or the same technique but employing different parameters and/or pre- or post-processing strategies (such as PLP and J-RASTA-PLP features) . Thus, the same (fullband) frequency domain undergoes different processing strategies leading to different feature representations which are used in individual recognizers, the errors of which are hoped to be complementary. The streams are recombined, just as in the MB approach according to FC processing, later in the process to dampen the errors.

5 Experimental Evaluations and Conclusions

The proposed algorithms for combining multiple subband or fullband streams, employing equal weighting, were tested on a continuously spoken digits database (Numbers95) [2] under noise-free (matched) conditions and under noise-corruption by artificial band-limited (stationary and siren) and natural wideband noise (mismatch) [8]. Results are presented for the case when HMM /MLP hybrid systems were used. As our goal was to develop systems which can easily generalize and adapt to unseen data, training was only carried out on clean speech. All tests were run using both PLP and J-RASTA-PLP features. Due to in general higher recognition rates in noisy speech, results are given for J-RASTA-PLP features only.

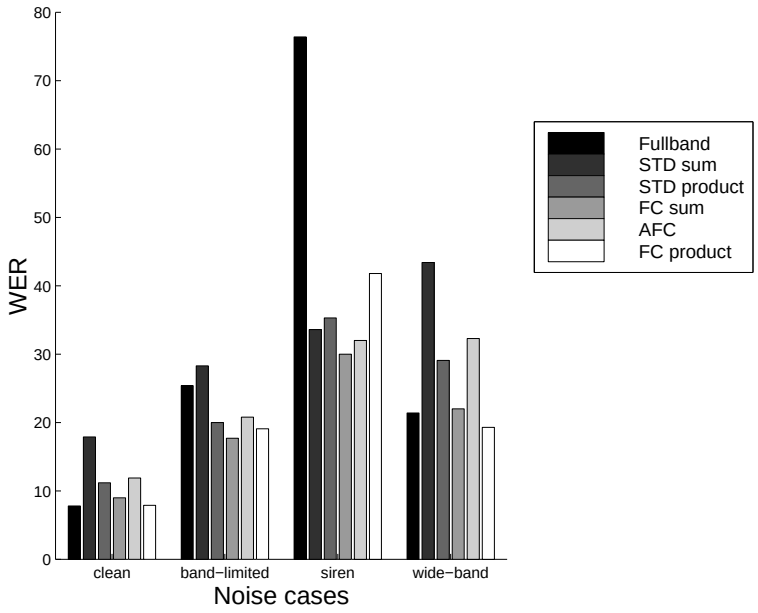

Fig. 2. Multi-band processing (4 subbands) using the "standard" STD (employing 4 trained MLPs) and "full combination" FC (employing 16 trained MLPs) recombination strategies as compared to the fullband approach on clean speech and speech corrupted with various additive noise cases. Features used are the j-rasta features.

It can be seen that the *MB FC* approach (FC Sum and FC Product) is competitive in clean speech (which is not the case for *standard* multi-band processing (STD sum and STD product)) and usually ranges among the best systems for all noise cases (cf. Figure 2). Depending on the noise case, it was observed that the FC Sum rule generally obtains better results in band-limited noise, and the FC Product in wideband noise.

For performance *improvement* in *clean* speech, *MS* processing should be applied (cf. Figure 3), though none of our systems tested gained a significant improvement over the best (i.e. J-RASTA-PLP) baseline.

For MS processing in noise, the results were less conclusive, but again it was observed that the FC SUM rule obtains better results in band-limited noise, whereas in wideband noise it is the FC PRODUCT rule.

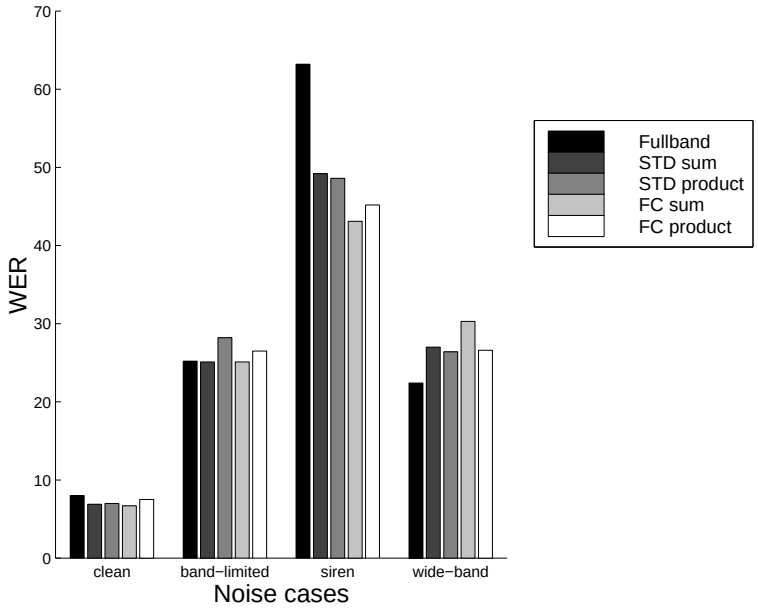

Fig. 3. Multi-stream processing using the "standard" STD (employing 3 trained MLPs) and "full combination" FC (employing 8 trained MLPs) recombination strategies as compared to the fullband (single-stream) approach on clean speech and speech corrupted with various additive noise cases. The streams consist of different features, namely PLP , J-RASTA-PLP and MFCC features. Fullband baseline is again the J-RASTA-PLP based system.

When looking at the performance achieved by PLP versus J-RASTA-PLP features, we observed that PLP features usually gain lower word error rate in clean and non-stationary band-limited noise whereas J-RASTA-PLP features performed better in stationary band-limited noise and wideband noise.

References

1. H. Christensen, B. Lindberg, and O. Andersen. Employing heterogeneous information in a multi-stream framework. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, III:1571–1574, 2000.

2. R.A. Cole, M. Noel, T. Lander, and T. Durham. New telephone speech corpora at CSLU. *Proc. European Conf. on Speech Communication and Technology*, 1:821–824, 1995.
3. M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable data. *Speech Communication*, 34(3):267–285, 2001.
4. M. Cooke, A. Morris, and P. Green. Missing data techniques for robust speech recognition. *Int. Conf. on Spoken Language Processing*, pages 863–866, 1997.
5. Stéfan Dupont. *Études et Développement de Nouveaux Paradigmes pour la Reconnaissance Robuste de la Parole*. PhD thesis, Laboratoire TCTS, Université de Mons, Belgium, 2000.
6. A.K. Haberstadt and J.R. Glass. Heterogeneous measurements and multiple classifiers for speech recognition. *Int. Conf. on Spoken Language Processing*, 3:995–998, 1998.
7. A. Hagen and H. Bourlard. Using multiple-time scales in the framework of full combination multi-stream speech recognition. *Int. Conf. on Spoken Language Processing*, 1:349–352, 2000.
8. A. Hagen and H. Glotin. Études comparatives des robustesses au bruit de l'approche 'full combination' et de son approximation. *Journée d'Études sur la Parole, Aussois*, pages 317–320, 2000.
9. Astrid Hagen. *Robust speech recognition based on multi-stream processing*. PhD thesis, Département d'informatique, École Polytechnique Fédérale de Lausanne, Switzerland, 2001.
10. K. Kirchhoff and J.A. Bilmes. Combination and joint training of acoustic classifiers for speech recognition. *ISCA ITRW Workshop on Automatic Speech Recognition – Challenges for the new millenium (ASRU2000)*, pages 17–23, 2000.
11. H. Meinedo and J.P. Neto. Combination of acoustic models in continuous speech recognition hybrid systems. *Int. Conf. on Spoken Language Processing*, 2:931–934, 2000.
12. A. Morris, M. Cooke, and P. Green. Some solutions to the missing features problem in data classification, with application to noise robust ASR. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 737–740, 1998.