Multiband with contaminated training data

Stéphane Dupont†‡, Christophe Ris†

†Faculté Polytechnique de Mons, TCTS Lab, Mons, Belgium
‡International Computer Science Institute, Berkeley, California, USA
{dupont,ris}@tcts.fpms.ac.be

Abstract

In this paper, we present a new approach for improving the robustness of automatic speech recognition systems to additive noise. This approach lies on the use of a particular training procedure (based on data contamination) in a particular architecture (the multi-band paradigm). In this framework, we expect to remove the drawbacks of both the corpus contamination approach which is the dependency to noise spectral characteristics, and the multi-band architecture which is its inefficiency in case of wideband noise. This method has been tested on the AURORA 2 task and compared to other robust methods such as spectral subtraction, J-RASTA filtering and missing data compensation, leading to very good performance on different kinds of additive noise, without any a priori knowledge of the noise characteristics.

1. Introduction

Additive noise is one of the most common sources of degradation of the performance of automatic speech recognition (ASR) systems. The effect of the noise is to lead to a mismatch between the acoustic models and the acoustic data. Various techniques have been developed in order to decrease the impact of noise on ASR systems, such as spectral subtraction [5, 6], J-RASTA filtering [4], model adaptation [14], missing data compensation [7, 8, 9], ...

One of the most efficient techniques to improve robustness of speech recognition systems on additive noise consists in training the acoustic models with data corrupted by noise at different signal-to-noise ratios (SNR) [3]. This approach leads to quasi-optimal performance when the noise used for training is spectrally similar to the noise used in the application but fails when the noises are spectrally too different. Therefore, this approach is useful if we have a good a priori knowledge on the noise spectrum characteristics. Another class of robust approaches is the sub-band analysis which consists in developing independent acoustic models in different frequency bands [10, 11]. It is therefore possible in a second step to weight the importance of those frequency bands according to their reliability and hence to minimize the influence of noisy bands. Unfortunately, this approach is still not particularly efficient in case of wideband noise.

The approach presented in this paper allows to get rid off the limitations of the last two robust techniques. Based on the multi-band architecture, our approach lies on the observation that, if we consider narrow frequency bands, noises inside the bands practically differ by their energy level only, not by the shape of their band limited power spectra. Therefore, we can train acoustic models associated with the multiple frequency bands on data corrupted by any kind of wideband noise at different signal-to-noise ratios. If the frequency bands are narrow enough, we can then expect these models to be robust to other kinds of noises. The bandwidth of the frequency bands (and consequently the number of subbands) will results from a trade-off between the assumption that noise is white within a subband and the ability to discriminate between speech and noise inside a subband.

So, the method consists essentially in the use of a particular training procedure (based on data contamination) in the framework of a particular architecture (based on sub-band analysis). As already stated, these two methods seem to have limited interest when used independently. Note also, that due to their complementarity with the proposed scheme, other robust methods such as spectral subtraction, filtering of the temporal feature trajectories, can also be combined in this architecture.

2. Description

This section describes our approach. We first perform a critical band analysis of the windowed speech frames. Similarly to PLP processing [2], this analysis uses a frequency domain filter-bank with 30 trapezoidal filters equally spaced along a Bark scale. The 30 critical band energies are then split into sub-vectors featuring the spectral envelope in different frequency bands. Among the different configuration that we have been testing, splitting in 7 bands, as shown in Figure 1, gave the best performance. Each sub-vector is then normalized in order to obtain parameters that are independent on the absolute energy of the speech frame. One could also compute subband cepstral coefficients by applying a discrete cosine transform on the sub-vectors. These two options were actually shown to give similar performance.

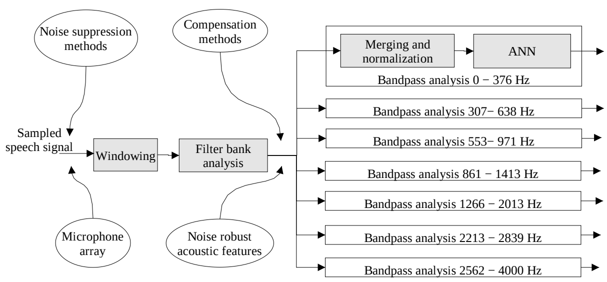

Figure 1: *Computation of robust acoustic features, related to 7 frequency bands.*

The key element of our approach is a scheme to estimate robust parameters from these subbands. To achieve this, each subband acoustic feature vector is non-linearly transformed using an artificial neural network (ANN). We actually use multi-layer perceptrons (MLPs) [1] designed for phonetic unit classification. As suggested in the introduction, training these MLPs on noisy data allow them to transform their input in an optimal way for noisy environments. Practically, white noise is added in a controlled way to the clean speech training corpus. As shown in Figure 2, this gives us a noisy training corpus with signal-to-noise ratios (SNRs) ranging from 0 dB to 20 dB, as well as a portion of clean speech. The 7 bands configuration that we have been using leads to frequency bands that are narrow enough to validate the 'white noise' assumption, while keeping enough speech specific information for phonetic classification within each band.

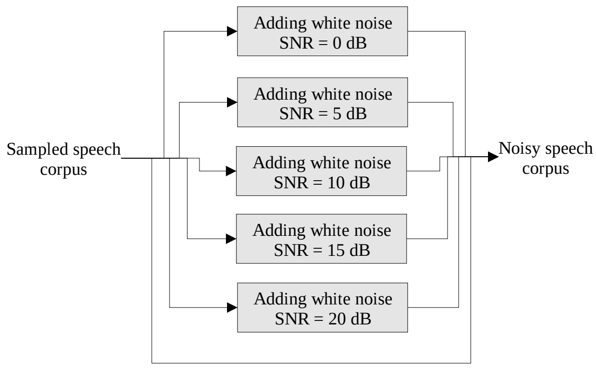

Figure 2: *Principle of training corpus contamination with white noise.*

In short, each subband uses an MLP trained to provide a nonlinear mapping between spectral acoustic features and phoneme posterior probability estimates. This mapping is optimized for phonetic classification in noisy environments. Single hidden-layer MLPs can provide probability estimates than can then be used as robust acoustic features for automatic speech recognition. To provide more flexibility, we rather used MLPs with two hidden layers (Figure 3). During recognition, the output of the second hidden layer is used as noise-robust acoustic feature vector for the corresponding subband. The size of the layer can be optimized or adjusted to get the desired number of features. This kind of approach is known as non-linear discriminant analysis (NLDA) [15]. A similar idea is also exploited in the Tandem speech recognition structure [12]. In our case however, multiple non-linear transformation are applied to obtain robust features into spectral subbands.

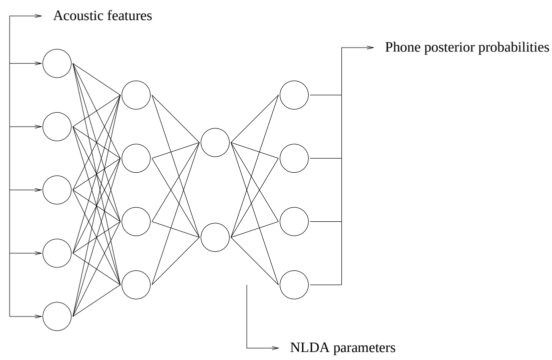

Figure 3: *Nonlinear discriminant analysis.*

The subband features are then concatenated to obtain an acoustic feature vector that can be used in any classical automatic speech recognition system. In our case, we have been using a HMM-based system with a MLP for acoustic modeling [1]. The multiband structure could be trained on different acoustic data and different phonetic units that the speech recognition acoustic model. These two systems can indeed be seen as two independent components: multiband robust feature extraction and speech recognition acoustic modeling. Training data with sufficient phonetic coverage might thus provide a multiband "feature extraction" structure that is portable across different tasks, and noise conditions. Preliminary results with a Tandem structure [13] even suggest portability across different languages.

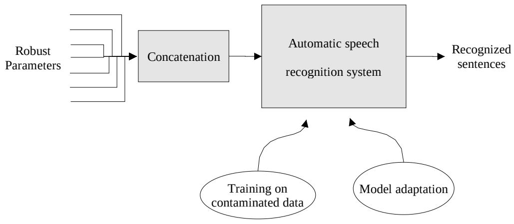

Figure 4: *Application to automatic speech recognition.*

Finally, this approach can easily be combined with complementary noise robust methods, for instance spectral subtraction and model adaptation (see Figures 1 and 4).

3. Experiments

Experiments have been carried out on the AURORA 2 [16] database. This database is based on TI-DIGITS (connected digits in american English) corrupted by different kinds of noises. We limited our experiments to the following four types of noise: subway, babble noise, in-car noise and exhibition hall.

The vocabulary is composed of the 10 digits, we have defined word models depicted by 127 HMM states.

The baseline hybrid system contains only 323,195 parameters. It uses 15 frames of 13 dimension acoustic features, it has 1000 hidden units and 127 outputs.

The multi-band system is composed of 7 subband MLPs with 2 hidden layers each. MLP are fed with 15 frames of frequency band specific spectral parameters. We have defined two configurations.

- The first one is quite heavy, multi-band MLPs have 1000 nodes in the first hidden layer, 30 nodes in the second hidden layer. The ASR system is a hybrid HMM/MLP modelizing the 127 HMM states. This MLP contains 1000 hidden nodes and is fed with 3 frames of concatenated vectors (that is $3 \times 7 \times 30 = 630$ input nodes). The global system contains 1,531,185 parameters.
- The second configuration aims at keeping the system as light as possible and to obtain a number of parameters in the same order than the baseline system. In this case, the multi-band MLPs have only 150 nodes in the first hidden layer. The ASR MLP, contains 500 hidden nodes and takes only one frame at its input (that is $7 \times 30 = 210$ input nodes). The total number of parameters for this configuration is 285,565.

	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB	clean	average 0→20 dB
Baseline (log-RASTA)	90.4%	68.4%	39.3%	19.2%	7.3%	3.1%	1.2%	27.5%
J-RASTA	82.9%	55.2%	27.5%	11.6%	4.8%	2.3%	0.9%	20.3%
Non-linear spectral subtraction	77.6%	50.0%	24.5%	10.6%	5.3%	3.2%	1.2%	18.7%
Missing data compensation	75.4%	43.7%	18.8%	7.1%	3.0%	1.7%	0.9%	14.9%
Contaminated multiband (conf.1)	59.5%	28.9%	11.9%	5.0%	2.4%	1.0%	0.5%	9.8%
Contaminated multiband (conf.2)	63.8%	33.5%	14.3%	6.3%	3.2%	1.7%	0.9%	11.8%
Matching noise conditions	54.3%	24.2%	8.6%	3.8%	2.0%	1.6%	1.3%	8.0%

Table 1: *Word error rate of different noise robust methods. Average on 4 kinds of noises for different SNR. Last column gives average WER for SNR from 0 dB to 20 dB (cf. AURORA official protocol)*

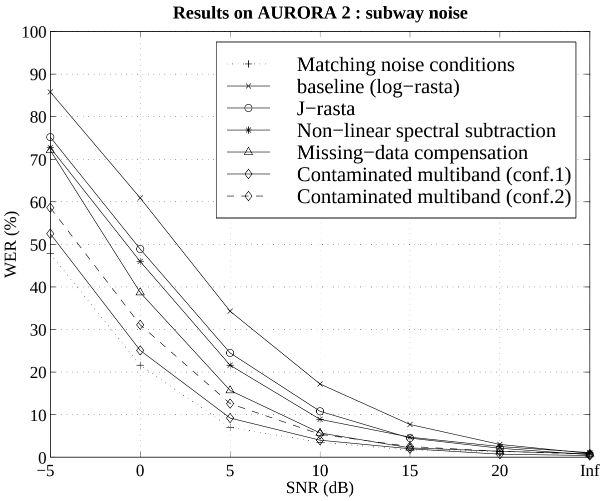

Figure 5: *Recognition performance of different robust techniques on subway noise.*

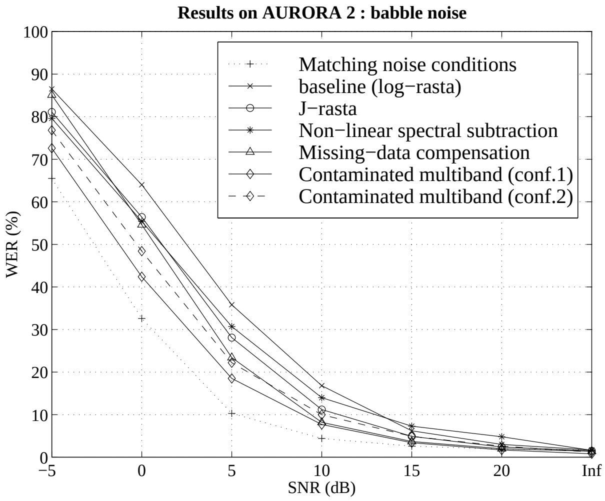

Figure 6: *Recognition performance of different robust techniques on babble noise.*

Note that, in our implementation, the multi-band MLPs and the ASR MLP have been trained on the same speech corpus, that is white-noise contaminated data.

Our system has been compared to:

- a baseline hybrid system trained on PLP derived from log-RASTA filtering of critical band energies,
- PLP derived from J-RASTA filtering [4] of critical band energies,
- PLP derived from non-linear spectral subtraction [6] of critical band energies,
- PLP derived from missing data compensated critical band energies [8, 9]. Practically we used a set of 256 gaussians to perform the missing data imputation. Selection of reliable spectral components was based on automatic local SNR estimation. See [7] for a complete theoretical description of the missing data techniques.
- as a reference, a J-RASTA PLP based hybrid system trained on data contaminated with the same noise as for the test (matching conditions).

Results for each noise are shown on Figure 5, 6, 7 and 8. Table 1 shows average results on the 4 kinds of noise.

As we can see in Table 1 our method outperforms other robust techniques and leads to relative improvements as high as 74% (at 10 dB SNR) compared to the baseline system and around 50% improvement compared to robust methods such as J-RASTA filtering or spectral subtraction. Note also that for some noises such as in-car noise for instance, improvement is even larger (up to 90% relative improvement compared to the baseline system - see Figure 7). Without any a priori assumption on the noise characteristics, we obtain recognition accuracy that gets close to the system trained on matching noise conditions. The proposed structure also outperforms the other systems in the case of clean speech, even though noise is used in the training procedure of the subband discriminant neural networks.

4. Conclusion

We have proposed a new algorithm for the optimal estimation of noise robust acoustic features. This estimation is based on the contamination of training data which is known to give quasi-optimal performance if the noise conditions are known a priori. In order to remove this constraint and to make the system independent to the noise spectral characteristics, we cut the signal into narrow frequency bands. In each subband, we can therefore assume that the noise is quasi-white justifying the training of subband MLPs on speech data contaminated by white noise at different levels. This MLPs can therefore be used for the estimation of acoustic features in each subband that can be assumed to be robust to any kind of noise.

Our approach has been tested on the AURORA task on 4 kinds of noise at SNR ranging from -5 dB to 20 dB and compared to other robust method described in the literature. We

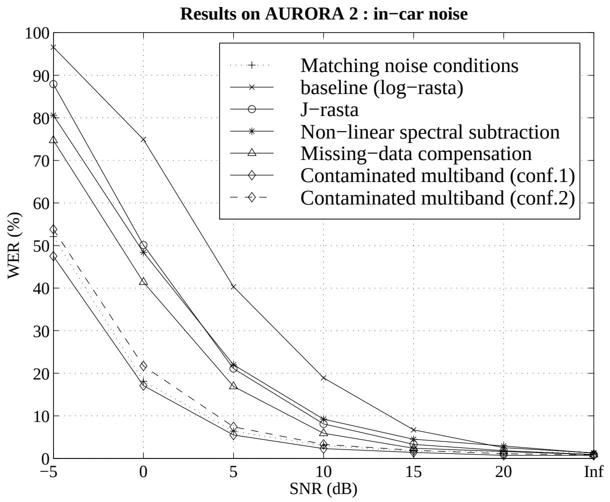

Figure 7: *Recognition performance of different robust techniques on in-car noise.*

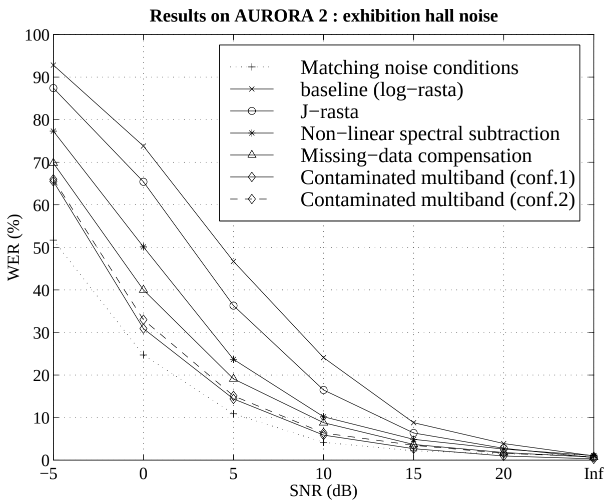

Figure 8: *Recognition performance of different robust techniques on exhibition hall noise.*

show that our method leads to an average (on different noises) relative diminution of the error rate up to 74% compared to our baseline system (log-RASTA PLP) and up to 57% compared to robust features such as J-RASTA PLP. And this without the need of any a priori knowledge on the noise characteristics. Additionally, the proposed system also yields improved performance in the case of clean speech.

Moreover, a particular attention was given to not increase the overall number of parameters of the ASR system in order to keep the system as competitive as possible.

As stated in the introduction, other noise robust techniques could be integrated to this architecture. More particularly, we think that the use of missing data compensated features as input to the multiband structure could further help.

5. References

[1] H. Bourlard and N. Morgan, “Connectionist Speech Recognition: A Hybrid Approach”, Kluwer, 1994.

[2] H. Hermansky, ”Perceptual linear predictive (PLP) analysis of speech”, The Journal of the Acoustical Society of America, vol.87, nr.4, april 1990, pp. 1738-1752

[3] T. Morii and H. Hoshimi, ”Noise Robustness in Speaker Independent Speech Recognition”, proceedings of ICSLP’90, pp. 1145-1148, 1990

[4] H. Hermansky and N. Morgan, ”RASTA processing of speech”, IEEE Trans. on Speech and Audio Processing, vol.2, nr.4, 1994, pp. 578-589

[5] M. Berouti, R. Schwartz and J. Makhoul, ”Enhancement of speech corrupted by acoustic noise”, proceedings of ICASSP’79, pp. 208-211, 1979

[6] P. Lockwood and J. Boudy, ”Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars”, Speech Communication, 11, pp. 215-228, 1992

[7] M. Cooke, P. Green, L. Josifovski and A. Vizinho, ”Robust automatic speech recognition with missing and unreliable acoustic data”, Speech Communication vol.34, no.3, june 2001, pp. 267-285

[8] M. Cooke, A. Morris and P. Green, ”Missing Data Techniques for Robust Speech Recognition”, proc. of ICASSP’97, Munich, 1997

[9] S. Dupont, ”Missing Data Reconstruction for Robust Automatic Speech Recognition in the Framework of Hybrid HMM/ANN Systems”, proc. ICSLP’98, Sydney, 1998

[10] S. Tibrewala and H. Hermansky, ”Sub-band based recognition of noisy speech”, proc. of ICASSP’97, Munich, 1997, pp. 1255-1258

[11] H. Bourlard, S. Dupont, H. Hermansky and N. Morgan, ”Towards sub-band-based speech recognition”, proc. of European Signal Processing Conference, Trieste, Italy, 1996, pp. 1579-1582

[12] H. Hermansky, D. Ellis and S. Sharma, ”Tandem connectionist feature extraction for conventional HMM systems”, proceedings of ICASSP’00, Istanbul, Turkey

[13] C. Bentez, L. Burget, B. Chen, S. Dupont, H. Garudadri, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan and S. Sivadas, ”Robust ASR front-end using spectral-based and discriminant features: experiments on the Aurora tasks”, proceedings of EUROSPEECH’01, Aalborg, Denmark, 2001

[14] C.J. Leggetter and P.C. Woodland, ”Maximum likelihood linear regression for speaker adaptation”, Computer Speech and Language, vol.9, pp. 171-185, 1995

[15] V. Fontaine, C. Ris and J.M. Boite, ”Nonlinear Discriminant Analysis for Improved Speech Recognition”, proceedings of EUROSPEECH’97, Rhodes, Greece, 1997.

[16] Aurora Project - Distributed Speech Recognition. Home Page: http://www.etsi.org/technicalactiv/dsr.htm