Multi-band with contaminated training data

Stéphane Dupont†‡, Christophe Ris†

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

Abstract

In this paper, we present a new approach for improving the robustness of automatic speech recognition systems to additive noise. This approach lies in the use of a particular training procedure (based on data contamination) in a particular architecture (the multi-band paradigm). With 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 relative inefficiency in the case of wideband noise. This method has been tested on the AURORA 2 continuous digits task and compared to other robust methods such as spectral subtraction, J-RASTA filtering and missing data compensation. It yields very good performance on different kinds of additive noise, without any a priori knowledge of the noise power spectrum.

1. Introduction

Additive noise is one of the most important sources of degradation of the performance of automatic speech recognition (ASR) systems. The effect of noise is to create 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 of the noise power spectrum. Another class of robust approaches is the sub-band analysis (or multi-band architecture) 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 the case of wideband noise.

The approach presented in this paper allows to get rid of the limitations of these two techniques. Based on the multi-band architecture, this approach follows from the observation that, if we consider narrow frequency bands, noises inside these 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 noise. The bandwidth of the frequency bands (and consequently the number of sub-bands) will results from a trade-off between the assumption that noise is white within a sub-band and the ability to discriminate between speech and noise, and between speech sounds, inside a sub-band.

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 rather limited interest when used independently. Note also, that due to their complementarity with the proposed scheme, other methods for speech recognition under noisy conditions (such as spectral subtraction or filtering of the temporal feature trajectories for instance) 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 configurations that we have been testing, a 7 bands split, as shown in Figure 1, gave the best performance. Each sub-vector is then normalized in order to obtain parameters that are independent of the absolute energy of the speech frame. One could also compute sub-band cepstral coefficients by applying a discrete cosine transform on the sub-vectors. These two options were actually shown to give similar performance.

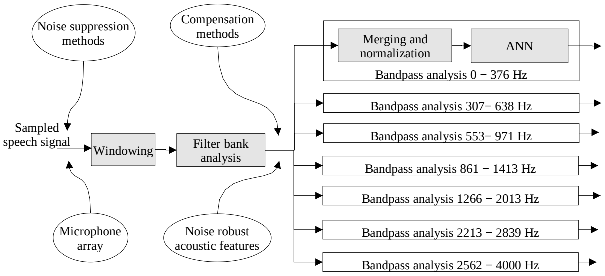
Noise suppression methods
Compensation methods
Merging and normalization
ANN
Bandpass analysis 0 − 376 Hz
Bandpass analysis 307− 638 Hz
Bandpass analysis 553− 971 Hz
Sampled speech signal
Windowing
Filter bank analysis
Bandpass analysis 861 − 1413 Hz
Bandpass analysis 1266 − 2013 Hz
Bandpass analysis 2213 − 2839 Hz
Bandpass analysis 2562 − 4000 Hz
Microphone array
Noise robust acoustic features

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 sub-bands. To achieve this, each sub-band acoustic feature vector is non-linearly transformed using an artificial neural network (ANN). We actually use feed-forward multilayer perceptrons (MLPs) [1] designed for phonetic unit classification. As suggested in the introduction, training these MLPs on noisy data allows 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.

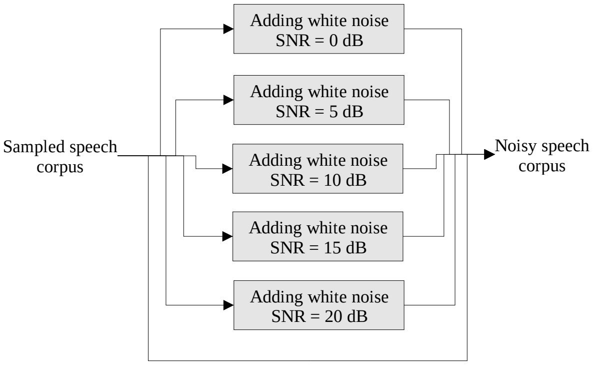
Sampled speech corpus
Adding white noise SNR = 0 dB
Adding white noise SNR = 5 dB
Adding white noise SNR = 10 dB
Adding white noise SNR = 15 dB
Adding white noise SNR = 20 dB
Noisy speech corpus

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

In short, each sub-band 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 acoustic feature vector for the corresponding sub-band. 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 transformations are applied to obtain robust features into spectral sub-bands. 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.

The sub-band 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 (or hybrid HMM/ANN system) [1]. The multi-band structure could be trained on different acoustic data and different phonetic units than the speech recognition acoustic model. These systems can indeed be seen as two independent components: multi-band robust feature extraction and speech recognition acoustic modeling. Training data with sufficient phonetic coverage might thus provide a multi-band ”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.

Finally, this approach can easily be combined with complementary noise robust methods, for instance spectral subtraction and model adaptation (see Figures 1 and 4). In our case, we will use temporal trajectories filtering.

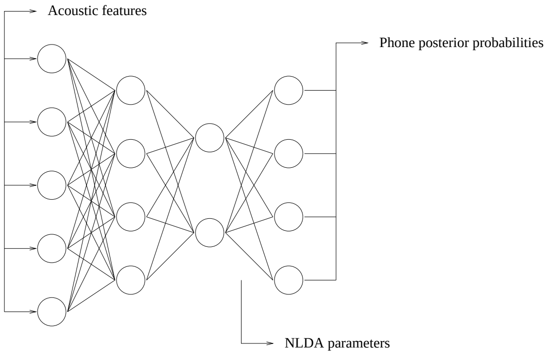

Figure 3: *Nonlinear discriminant analysis.*

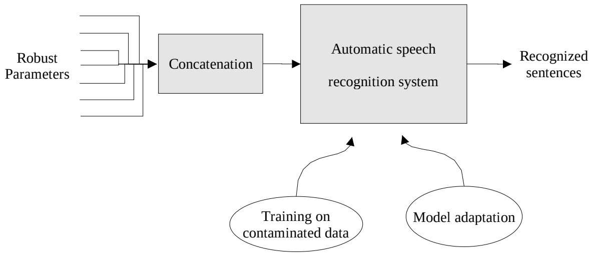

Figure 4: *Application to automatic speech recognition.*

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 noise. 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, and the back-end uses word models depicted by a total of 127 HMM states.

For comparison purposes, we have been using systems based on different kinds of acoustic feature:

- PLPs derived from log-RASTA filtering of critical band energies,
- PLP derived from J-RASTA filtering [4] (J parameter fixed to 10^{-6}) of critical band energies,
- PLP derived from non-linear spectral subtraction [6] applied to 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, we also used a J-RASTA PLP based system trained on data contaminated with the same noises as the test data (matching train/test noise).

	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB	clean	average 0→20 dB
log-RASTA PLP	90.4%	68.4%	39.3%	19.2%	7.3%	3.1%	1.2%	27.5%
J-RASTA PLP	82.9%	55.2%	27.5%	11.6%	4.8%	2.3%	0.9%	20.3%
Non-linear spectral subtr. PLP	77.6%	50.0%	24.5%	10.6%	5.3%	3.2%	1.2%	18.7%
Missing data compensation PLP	75.4%	43.7%	18.8%	7.1%	3.0%	1.7%	0.9%	14.9%
J-RASTA contaminated multi-band (conf.1)	59.5%	28.9%	11.9%	5.0%	2.4%	1.0%	0.5%	9.8%
J-RASTA contaminated multi-band (conf.2)	63.8%	33.5%	14.3%	6.3%	3.2%	1.7%	0.9%	11.8%
Matching train/test noise	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 noise for different SNR. Last column gives average WER for SNR from 0 dB to 20 dB (cf. AURORA official protocol)*

Each of these baseline systems uses a MLP for acoustic modeling (hybrid HMM/ANN approach). These MLPs have 15 frames of 13 dimension acoustic features (energy + 12 PLPs) as input, 1000 hidden units and 127 outputs. They have 323,195 parameters.

The multi-band system is composed of 7 sub-band MLPs with 2 hidden layers each. They use 15 frames of frequency band specific spectral parameters. J-RASTA temporal trajectory filtering is applied to the outputs of the critical band filter-bank (J parameter fixed to 10^{-6}) before feeding these MLPs. Two configurations have been defined:

- 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 modeling the 127 HMM states. This MLP contains 1000 hidden nodes and uses 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.

Note that, in our implementation, the sub-band MLPs and the ASR MLP have been trained on the same white-noise contaminated TI-DIGITS training corpus.

Results are shown in Figures 5, 6, 7 and 8. Table 1 shows average results for the four noises.

As we can see, our method outperforms other robust techniques and leads to a relative average error rate reduction of 64% compared to the baseline system and of 50% compared to robust methods such spectral subtraction or J-RASTA filtering used alone. Note also that for band-limited noises such as the in-car noise, improvement is even larger (up to 90% relative improvement compared to the baseline system - see Figure 7). Without any knowledge of the noise statistics, we obtain recognition accuracy that gets close to the accuracy of 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 sub-band discriminant neural networks.

4. Conclusions

We 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 operating noise conditions are known a priori. In order to get rid of this constraint, and to make the system independent of the noise spectral characteristics, we cut the signal into narrow frequency bands. In each sub-band, we can therefore assume that the noise is quasi-white justifying the training of sub-band MLPs on speech data contaminated with white noise at different signal-to-noise ratios. These MLPs can therefore be used to estimate acoustic features in each sub-band. These features can be assumed to be robust to any kind of noise.

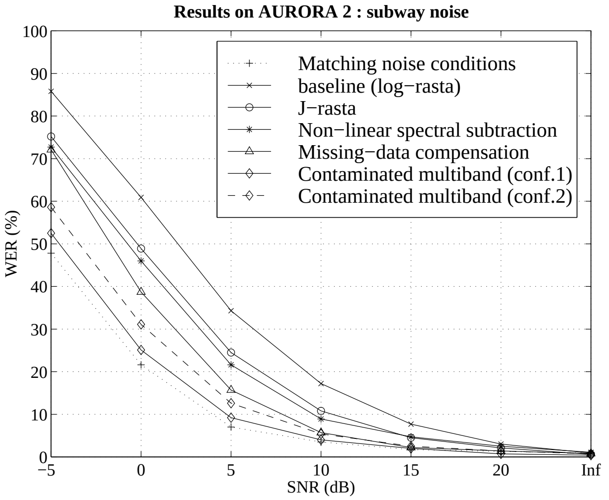

Figure 5: *Recognition performance for different robust techniques on subway noise.*

Our approach has been tested with the AURORA 2 task on 4 kinds of noise at SNRs ranging from -5 dB to 20 dB, and compared to other robust methods described in the literature. We showed that our method leads to a relative reduction of the average (over different noises) error rate of 50% compared to robust features such as J-RASTA PLPs. This gain is obtained 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 avoid increasing the overall number of parameters of the ASR system in order to keep it as competitive as possible.

Other noise robust techniques could be integrated to this architecture. Although we already tried J-RASTA filtering, we think that the use of missing data compensated features as input to the multi-band structure could further help.

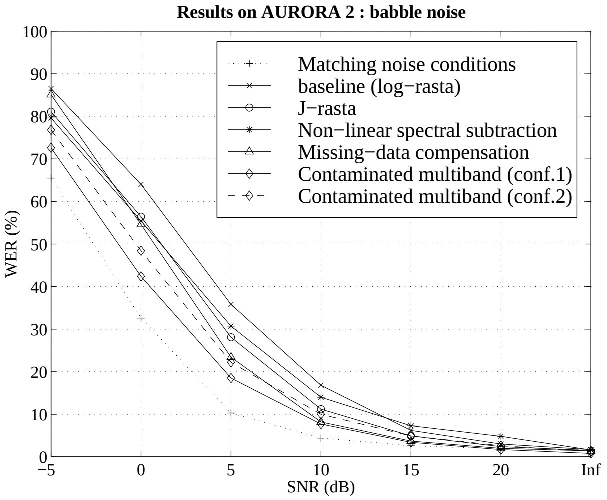

Figure 6: *Recognition performance for different robust techniques on babble noise.*

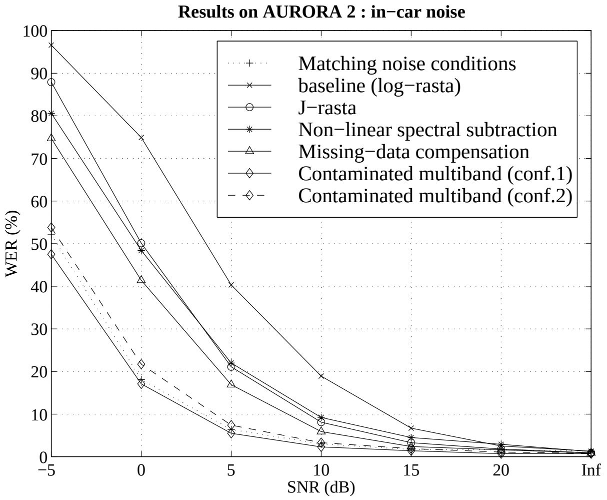

Figure 7: *Recognition performance for different robust techniques on in-car noise.*

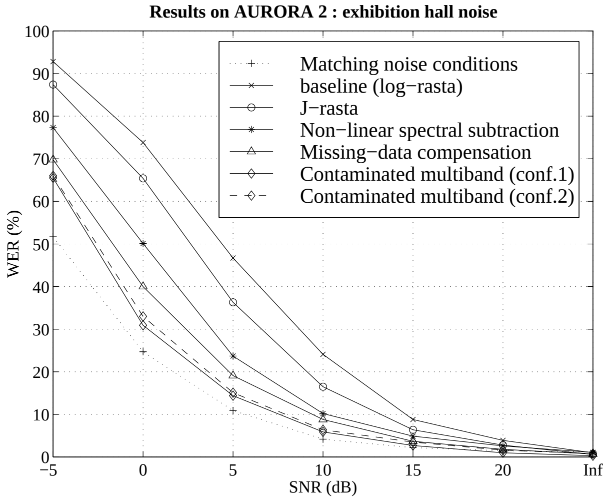

Figure 8: *Recognition performance for different robust techniques on exhibition hall noise.*

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. Benítez, 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