Robust feature extraction and acoustic modeling at Multitel: experiments on the Aurora databases

Stéphane Dupont and Christophe Ris[1]

Multitel & FPMS-TCTS, Avenue Copernic 1, B-7000 Mons, Belgium

dupont,ris@multitel.be

Abstract

This paper intends to summarize some of the robust feature extraction and acoustic modeling technologies used at Multitel, together with their assessment on some of the ETSI Aurora reference tasks. Ongoing work and directions for further research are also presented.

For feature extraction (FE), we are using PLP coefficients. Additive and convolutional noise are addressed using a cascade of spectral subtraction and temporal trajectory filtering. For acoustic modeling (AM), artificial neural networks (ANNs) are used for estimating the HMM state probabilities. At the junction of FE and AM, the multi-band structure provides a way to address the needs of robustness by targeting both processing levels. Robust features within sub-bands can be extracted using a form of discriminant analysis. In this work, this is obtained using sub-band ANN acoustic models. The robust sub-band features are then used for the estimation of state probabilities.

These systems are evaluated on the Aurora tasks in comparison to the existing ETSI features. Our baseline system has similar performance than the ETSI advanced features coupled with the HTK back-end. On the Aurora 3 tasks, the multi-band system outperforms the best ETSI results with an average reduction of the word error rate of about 62% with respect to the baseline ETSI system and of about 18% with respect to the advanced ETSI system. This confirm previous positive experience with the multi-band architecture on other databases.

1. Introduction

Robustness to speech signal degradations such as environmental noise or channel distortions is currently one of the main challenges in automatic speech recognition, especially, when more and more commercial applications are emerging, confronting the technology to practical usage conditions.

In this paper, we introduce some of the last advances in robust feature extraction and acoustic modeling achieved at Multitel and their assessment on the ETSI Aurora 3 reference task. More particularly, we present a system based on a multi-band architecture which seems to constantly improve the robustness to additive noise on different speech recognition tasks. This approach is taking advantage of the independent processing of narrow frequency bands to efficiently handle the mismatch between training and testing conditions through a noise "vaccination" technique yielding robustness to noises not seen during the training.

In section 2, we introduce a technical description of our full-band and multi-band systems in terms of feature extraction and acoustic modeling. In this section, we also present former results on different ASR tasks. In section 3, we compare both the full-band and the multi-band systems to the ETSI features [16] on the Aurora 3 reference task. In section 4, we discuss the results and propose some ideas for future improvements.

[1]This work is sponsored by the Walloon Region through the "Initiative" MODIVOC project.

2. Feature Extraction and Acoustic Modeling

This section presents the feature extraction and acoustic modeling approaches that have been used to handle noise and channel distortions. Although the evaluations on the Aurora tasks are part of Section 3, results on other databases will be recalled to provide a broader view and more specifically to confirm the consistency of the improvements that we can get using the multi-band structure.

2.1. Feature extraction

The reference front-end is based on PLP processing [13] but MFCC processing yields similar recognition performance. Additive and channel noises are handled using a combination of spectral subtraction (SPS) and temporal trajectory filtering. The spectral subtraction is applied on the FFT frequency bins. It has been shown in the past [9] that smoothing of the SPS filter is important to reduce the amplitude of artifacts (musical noise) and provide some performance benefit at low signal-to-noise ratios (SNRs). This has been confirmed in our experiments (not reported here). Also, we have been using a linear-phase smoothing of the SPS filter. This was shown to yield better results than the non-linear-phase smoothing proposed in [1].

Right after spectral subtraction, critical band energies are computed. A band-pass filtering is then applied on the time trajectories of the different critical bands. Here again, we have observed that a linear-phase filter significantly outperforms the "traditional" RASTA filter [14], confirming the results published in [6].

2.2. Full-band acoustic modeling

Our speech recognition systems are based on the hybrid HMM/ANN architecture [4] where ANNs - typically multilayer perceptrons (MLP) - are used to estimate the HMM state likelihoods. This collaboration between ANNs and HMMs has proven its efficiency on many different speech recognition tasks, ANNs providing powerful locally discriminant acoustic modeling, and simplifying the integration of contextual information.

In a classical use, denoted "full-band acoustic modeling" in this paper, a single ANN is trained in order to classify the frames of acoustic features into language dependent acoustic units (phonemes, diphones, ...).

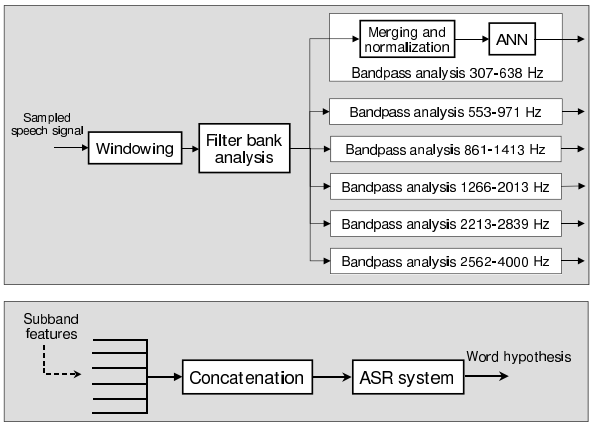

Figure 1: *Multitel multi-band system.*

2.3. Multi-band NLDA and acoustic modeling

In this approach, we propose to use a multi-band architecture [3, 12] in order to extract noise robust acoustic features. The principle 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 apply some discriminant analysis within each frequency band on data corrupted by any kind of wideband noise at different signal-to-noise ratios. If the frequency bands are narrow enough, we can expect the recombined acoustic features to be robust to other kinds of noises.

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 a multi-layer perceptron (MLP). 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. In our configuration we have defined 6 sub-bands (depicted in fig.1). For each sub-band, a MLP is 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.

In order to keep more flexibility, we used MLPs with two hidden layers. During recognition, the output of the second hidden layer is used as noise-robust acoustic feature vector for the corresponding sub-band. The size of this 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) [11]. The sub-band features are then concatenated to obtain an acoustic feature vector that can be used in any classical automatic speech recognition system (cf. fig. 1), a hybrid HMM/ANN system, in our case.

Training data with sufficient phonetic coverage might provide a multi-band ”feature extraction” structure that is portable across different tasks, noise conditions, or even different languages.

The resulting acoustic model is robust to additive noise. This is the case even if the noise strongly differs from the noise seen during the training process in both time and frequency structure.

2.4. Prior experimental results

Earlier results have been published in [7]. The goal of this section is to briefly recall these results so as to present a broader view of the multi-band vs. full-band performance. Together with the new experiments of Section 3, they demonstrate the consistent improvement that our multi-band structure can bring on different tasks (ranging from 10 digits to 1000 words), different languages, different noises and speaking styles (read speech vs. natural speech).

The results that were published in [7] are recalled in Table 1. Two databases have been used:

OGI Numbers [5] [20]: these are sequences of numbers recorded from the US landline telephone network. To make the task more challenging with respect to the robustness to noise, different noises have been added to the original data at different signal to noise ratios, ranging from 5 dB to 20 dB. Six noises types have been used: Gaussian white noise, ’lynx’ helicopter noise from Noisex corpus, car noise from the Madras corpus, noise inside a car, noise in a public hall and noise from a shopping mall.

Resource Management: we have been using the RM1 speaker independent portion of the data, the official train/test definition and the February 89 test set. The dictionary contains 992 words and the standard word-pair language model with a perplexity of 60 has been used. The original test data has been corrupted using the ’Lynx’ helicopter noise from the Noisex-92 database. This allowed comparison with results from the literature [18].

Task & SNR	FB WER (%)	MB WER (%)	WERR (%)
NB 20dB	10.0%	7.5%	25%
NB 10dB	15.4%	10.9%	29%
NB 5dB	23.0%	16.9%	27%
RM 20dB	28.2%	10.9%	61%
RM 12dB	63.0%	35.3%	44%

Table 1: *Word error rate of full-band (FB) and multi-band (MB) systems, and relative word error rate reduction (WERR) of the MB over the FB. ”NB” is the continuous numbers recognition task (OGI Numbers) and ”RM” is the Resource Management 1000 words continuous speech recognition task.*

PLP feature extraction has been used for these experiments and we only applied a RASTA temporal trajectory filtering to obtain some resistance to noise and channel effects (spectral subtraction was not applied for these experiments). As can be observed in the Table, the multi-band structure yields further increased robustness and a relative reduction of the error rate of more than 25% on the OGI Numbers task and of more than 50% on the Resource Management task. On this task, our system also presents a 25% reduction of the error rate with respect to results published in [18] using a model-based speech enhancement technique.

3. Experiments on Aurora 2 and Aurora 3

Experiments on the Aurora 2 database (English digits) have been conducted and reported in [8]. Briefly, with the multi-band structure, we have been achieving a 9.8% word error rate (in comparison with 41.3% for the ETSI baseline and 12.4% for the ETSI advanced front-ends with HTK) on test set A and the clean training data.

For this paper, experiments were then performed on the set of connected digit tasks that cover the four different languages of the Aurora 3 database: Finnish (FI), Spanish (SP), German (GE), and Danish (DA). These tasks are taken from the corpora that were recorded as part of the SpeechDat-Car European (SDC) project [21]. These are real recordings made in cars with a setup consisting of a close talking microphone and a distant microphone. Three train and test configurations were defined: the well-matched condition (WM), the medium mismatched (MM) condition and the highly mismatched condition (HM). In the WM case, 70% of the entire data is used for training and 30% for testing. The training set contains all the variability that appears in the test set. In the MM case, only far microphone data is used for both training and testing. For the HM case, training data consists of close microphone recordings only while testing is done on far microphone data.

3.1. Results

The configuration of the recognizer is fixed for all tasks. Each word is modeled by an HMM composed of 16 states. The proposed systems are evaluated in comparison to the existing ETSI DSR "regular" features [10] as well as advanced features [19], coupled with the reference HTK acoustic modeling and back-end used for the ETSI evaluations.

Results are presented in Tables 2, 3, 4 for the following systems:

- the ETSI regular (MFCCs) and advanced (robust) features coupled with the HTK back-end (results provided by the Aurora consortium and taken from [19]),
- the Multitel robust full-band baseline system,
- the Multitel robust multi-band system,

In the Tables, the relative improvements are computed as relative reductions in WER with respect to the ETSI regular system. Also, in computing the *Overall* performance for each task, the performance of *WM, MM and HM* conditions are weighted by *0.40, 0.35 and 0.25* respectively.

It can be seen that the reduction in WER achieved through the use of the multi-band structure is significant. For the Finnish language, the error rate is divided by a factor of 2 with respect to the ETSI advanced system. Consistency across the different languages/experiments is not perfect however and the Danish tasks seems problematic. Further investigations are under way. Viterbi alignement problems during the training are suspected. Also, no frame dropping procedure has been applied at this stage.

Aurora 3 Referance Word Error Rate					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	7.26%	7.06%	8.80%	12.72%	8.96%
Mid (x35%)	19.49%	16.69%	18.96%	32.68%	21.96%
High (x25%)	59.47%	48.45%	26.83%	60.63%	48.85%
Overall	24.59%	20.78%	16.86%	31.68%	23.48%

Aurora 3 Word Error Rate					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	3.91%	3.36%	4.89%	6.63%	4.70%
Mid (x35%)	19.08%	6.08%	9.16%	18.51%	13.21%
High (x25%)	13.39%	8.45%	8.75%	20.41%	12.75%
Overall	11.59%	5.58%	7.35%	14.23%	9.69%
Rel. Improv.	38.56%	63.85%	52.71%	50.91%	51.51%

Table 2: *Aurora 3 - ETSI regular (MFCCs) and ETSI advanced front-end coupled with ETSI HTK-based back-end.*

Aurora 3 Word Error Rate					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	3.10%	2.30%	7.00%	6.90%	4.82%
Mid (x35%)	9.10%	4.70%	14.70%	19.50%	12.00%
High (x25%)	12.70%	10.40%	14.90%	22.50%	15.13%
Overall	7.60%	5.17%	11.67%	15.21%	9.91%

Aurora 3 Relative Improvement					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	57.30%	67.42%	20.45%	45.75%	47.73%
Mid (x35%)	53.31%	71.84%	22.47%	40.33%	46.99%
High (x25%)	78.64%	78.53%	44.47%	62.89%	66.13%
Overall	61.24%	71.75%	27.16%	48.14%	52.07%

Table 3: *Aurora 3 - Multitel full-band system.*

Aurora 3 Word Error Rate					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	2.40%	2.00%	4.70%	6.00%	3.78%
Mid (x35%)	7.00%	4.40%	9.40%	20.80%	10.40%
High (x25%)	8.40%	5.90%	9.90%	20.70%	11.23%
Overall	5.51%	3.82%	7.65%	14.86%	7.96%

Aurora 3 Relative Improvement					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	66.94%	71.67%	46.59%	52.83%	59.51%
Mid (x35%)	64.08%	73.64%	50.42%	36.35%	56.12%
High (x25%)	85.88%	87.82%	63.10%	65.86%	75.66%
Overall	70.68%	76.40%	52.06%	50.32%	62.36%

Table 4: *Aurora 3 - Multitel multi-band system.*

4. Discussion - Future work

4.1. The multi-band architecture

A possible drawback from a multi-band structure comes from the fact that inter-band feature correlations are not appropriately taken into account by the within band discriminant analysis of the first processing stage (top of Figure 1). In general, this can potentially affect classification rate in the case the information in each band does not allow proper classification whereas using the information from both bands would. In practice however, both frame classification and ASR experiments show that the multi-band structure that we propose does not hurt performance on high SNR or clean speech. This has been systematically confirmed on different tasks and settings, including both context independent and context dependent acoustic modeling. In some cases, the multi-band structure even improves ASR performance on clean speech. Also, the "feature correlation" drawback is possibly compensated by strong robustness to (in addition to environmental noise) any frequency dependent factor including channel effects or variations of the speech signal which can be intrinsic to the speech production itself or to the coupling with the microphone (breath noise, microphone proximity effect...).

It is important to emphasize that the multi-band structure can probably be cascaded or inserted in any ASR architecture. The use of ANNs to get discrimination within the sub-bands allow a lot of flexibility in the choice of input features, even with strong correlation between the feature vector elements. Also, the discriminant features or the output of the "recombination" network could be used as input to other kinds of acoustic models (for instance mixtures of Gaussians) through a structure know as "Tandem" processing [15]. Both aspects will be investigated further (see improved sub-band modeling and generic feature extraction in the next section).

4.2. Ongoing and future work

The multi-band system has scarcely been developed further in this work as the algorithms and implementations that were used are almost exactly the one used in previous publications [8]. Further work is however going on in the following directions and convincing outcomes will be presented at the conference:

Improved sub-band modeling: using multiple streams of different features and ensemble of statistical models has consistently been shown to provide improvements of ASR performance [15, 17]. In the multi-band structure, different feature sets representing sub-band information will be used to develop ensembles of sub-band ANNs. Appropriate recombination of the sub-band discriminant features is then expected to further improve robustness and classification within each sub-band, and hence globally.

Improved multi-band modeling: in the current situation, the multi-band structure is composed of sub-band covering a group of critical bands. Also, no overlap between the sub-bands exists, except for the small overlap due to the shape of critical band filters from the PLP analysis. We are currently investigating a structure where the sub-bands overlap. This can yield improved ensemble properties through an increase in the number of ANNs. Also, as the width of each sub-band is kept the same as in the "reference" structure, no drawback is expected in terms of robustness. In addition to averaging properties due to the ensemble processing, there are psychoacoustic motivations for using overlapping bands.

Generic feature extraction: in the current study, the sub-band discriminant ANNs are developed on the task/language specific training corpora. We will be investigating the development of task/language independent sub-band discriminants. We will have to pool available data and define appropriate targets for the optimization of the discriminant ANNs.

Multi-band DSR and fixed point: a DSR version of the multi-band structure will be investigated. This will include an appropriate feature representation compression for transmission from the client to the DSR server and fixed point processing. This will be prototyped within the Modivoc [2] DSR system which already provides DSR using the full-band structure, with fixed point processing available for the whole ASR engine (FE, AM and continuous speech decoder).

LVCSR: the application of the multi-band structure (and the evaluation on the Aurora4 Wall Street Journal data) will imply the investigation and choice of appropriate targets for training the sub-band systems. Large vocabulary generally require the use of context dependent phonetic units.

5. Conclusions

This paper presents a summary and an evaluation of some of the latest robust feature extraction (FE) and acoustic modeling (AM) techniques used at Multitel.

First, it has been confirmed in this work that a linear phase filter outperforms the "traditional" RASTA filter. Then, the multi-band FE/AM structure has been presented. By presenting new results on the Aurora tasks and recalling earlier results (Section 3.1), this paper confirms the robustness and consistency of the multi-band structure on a broad range of tasks. On the Aurora3 tasks, the word error rate of the multi-band system is 62% lower than the error rate of the ETSI baseline. This has been achieved without frame dropping. Also, these results have been obtained without any particular tuning of the modeling approaches.

6. References

[1] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan and S. Sivadas, "Qualcomm-ICSI-OGI Features for ASR", in Proc. of Internat. Conf. on Spoken Language Processing, Denver, 2002.

[2] M. Bagein, O. Pietquin, C. Ris, G. Wilfart, "Enabling speech based access to information management systems over wireless networks", proceedings of ASWN 2003, Berne, July 2030

[3] 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

[4] H. Bourlard and N. Morgan, "Connectionist Speech Recognition: A Hybrid Approach", Kluwer, 1994.

[5] R. Cole et al., "Telephone speech corpus at CSLU", in Proc. of International Conference on Spoken Language Processing, Yokohama, Japan, September 1994.

[6] J. de Veth and L. Boves, "Effectiveness of phase-corrected RASTA for continuous speech recognition Information", in Proceedings Internat. Conf. on Spoken Language Processing, 1998

[7] S. Dupont, "Etude et Développement d'Architectures Multi-Bandes et Multi-Modales pour la Reconnaissance Robuste de la Parole", PhD thesis, Faculté Polytechnique de Mons, 2000.

[8] S. Dupont and C. Ris, "Multi-band with contaminated training data", in Proc. of CRAC workshop on consistent and reliable acoustic cues for sound analysis, Aalborg, Denmark, Sept. 2001.

[9] Y. Ephraim, "Statistical-model-based speech enhancement systems", Proceedings of the IEEE, 80(10), October 1992.

[10] *ETSI ES 201 108 v1.1.2 Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms,* April 2000.

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

[12] A. Hagen and A.C. Morris, "Recent advances in the multi-stream HMM/ANN hybrid approach to noise robust ASR", Computer, Speech and Language, 2002, submitted paper.

[13] 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

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

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

[16] H. G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", ISCA ITRW ASR2000 on Automatic Speech Recognition: Challenges for the Next Millennium, Paris, France, September 18-20, 2000.

[17] M. Kleinschmidt and D. Gelbart, "Improving Word Accuracy with Gabor Feature Extraction", in Proc. of International Conference on Spoken Language Processing, Denver, 2002.

[18] B. Logan, "Adaptive Model-Based Speech Enhancement", PhD thesis, University of Cambridge, 1998.

[19] D. Macho, L. Mauuary, B. Noe, Y. M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce and F. Saadoun, "Evaluation of a noise-robust DSR Front-End on Aurora Databases", in Proc. of Internat. Conf. on Spoken Language Processing, Denver, 2002.

[20] N. Mirghafori and N. Morgan, "Combining Connectionist multi-band and full-band probability streams for speech recognition of natural numbers", in Proc. of International Conference on Spoken Language Processing, Sydney, Australia, December 1998.

[21] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, and J. Allen., *Speechdat-car a large speech database for automotive environments,* In LREC (Language Resources and Evaluation Conference), Athens, 2000.