INTERSPEECH 2009 BRIGHTON

Hierarchical Processing of the Modulation Spectrum for GALE Mandarin LVCSR system

Fabio Valente[1], *Mathew Magimai.-Doss*[1], *Christian Plahl*[2], *Suman Ravuri*[3]

[1]IDIAP Research Institute, CH-1920 Martigny, Switzerland
[2] Human Language Technology and Pattern Recognition
Computer Science Department, RWTH Aachen University, Germany
[3] International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704

{fabio.valente,mathew}@idiap.ch, plahl@i6.informatik.rwth-aachen.de,ravuri@icsi.berkeley.edu

10.21437/Interspeech.2009-750

Abstract

This paper aims at investigating the use of TANDEM features based on hierarchical processing of the modulation spectrum. The study is done in the framework of the GALE project for recognition of Mandarin Broadcast data. We describe the improvements obtained using the hierarchical processing and the addition of features like pitch and short-term critical band energy. Results are consistent with previous findings on a different LVCSR task suggesting that the proposed technique is effective and robust across several conditions. Furthermore we describe integration into RWTH GALE LVCSR system trained on 1600 hours of Mandarin data and present progress across the GALE 2007 and GALE 2008 RWTH systems resulting in approximatively 20% CER reduction on several data set.

Index Terms: TANDEM features, speech recognition.

1. Introduction

TANDEM feature extraction [1] represents an effective method for transforming phoneme posterior distributions into features for conventional speech recognition systems based on Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM). Typically phoneme posteriors are obtained from a Multi Layer Perceptron (MLP) trained in order to discriminate between phonetic targets as described in [2]. A large number of studies (see e.g. [3]) have reported considerable improvements in large vocabulary speech recognition systems using the TANDEM based features. Furthermore they show complementary properties w.r.t. conventional short term feature extraction techniques.

Previously proposed MLP inputs are based on short term spectrum features e.g. MFCC, PLP or the modulation spectrum of the speech signal (i.e. long segments of spectral energy trajectories obtained by Short Term Fourier Transform [4],[5]). In our previous related work [6], we presented a technique that process separate ranges of *modulation frequencies* using independent classifiers. This approach is somehow inspired from the conventional multi-band approach. While multi-band operates on different ranges of auditory frequencies, the proposed method operates on different ranges of modulation frequencies. In [6], we carried the investigation on a large vocabulary task for transcription of meeting recordings acquired in several acoustic conditions. Experiments revealed considerable improvement when the available range of modulation frequencies is split in two parts and processed by independent classifiers. Furthermore we verified that the combination of classifiers trained on separate modulation frequencies sub-bands is more effective if performed in hierarchical (sequential) fashion rather then in parallel fashion (as in conventional multi-band ASR).

In this work, we first study in detail the approach of processing separate modulation frequency ranges (see [6]) on Mandarin Broadcast recordings. We aim at investigating if the observations made on the meeting data generalizes to other type of data as well (i.e. clean condition). We also investigate the relative improvements obtained by augmenting the modulation spectrum representation with other features: the pitch estimates (already proven effective in ASR for Mandarin language; for details see [7]) and the use of short term log critical band energy. Those experiments are run on the SRI/UW/ICSI system using 100 hours of training data.

In the second part, we describe the integration of the best TANDEM feature (resulting from hierarchical processing augmented with pitch and critical band energy) into RWTH GALE LVCSR system trained on 1600 hours of speech. Evaluation studies are conducted on four different data set (eval06, dev07, eval07 and dev08) and contrastive results are reported for the GALE 2007 and 2008 RWTH evaluation systems.

The 100 hours setup shows that hierarchical processing yields the best performance. The feature augmentation with pitch and log critical band energy can further reduce the system error. The experiments on RWTH GALE LVCSR system reconfirms the latter observation on large training data setup.

The reminder of the paper is organized as follows: Section 2 describes the LVCSR system for 100 hours Mandarin SRI/UW/ICSI system for the GALE project, Section 2.1 describes the principle of the hierarchical modulation spectrum and preliminary results on the eval06 data set, Section 2.2 describes the integration of pitch and critical band energy. Section 3 describes experiments on RWTH 1600 hours system and finally section 4 conclude the paper.

2. Investigations on 100 hours data setup

The following preliminary experiments are based on the large vocabulary ASR system for transcription of Mandarin language described in [8], developed by SRI/UW/ICSI for the GALE project. Recognition is performed using the SRI Decipher recognizer and results are reported in terms of Character Error Rate (CER). The training is done using approximatively 100 hours of Mandarin Broadcast data consisting of equal amount of Broadcast News and Broadcast Conversation data. The results are reported on the DARPA GALE evaluation 2006 data, i.e., eval06 data set. The baseline system uses 13 standard mel-frequency

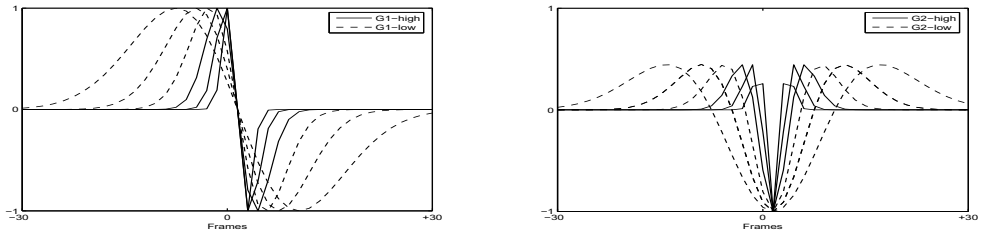

Figure 1: Set of temporal filter obtained by first (G1 left picture) and second (G2 right picture) order derivation of Gaussian function. G1 and G2 are successively split in two filter bank (F1 = {G1-low and G2-low}, dashed line) and (F2 = {G1-high and G2-high} continuous line) that filter respectively high and low modulation frequencies.

cepstral coefficients (MFCC) plus first and second order temporal derivatives. Mandarin is a tonal language thus the MFCC vector is augmented with smoothed log-pitch estimate (referred as f0) plus its first and second order temporal derivatives as described in [7] resulting in an acoustic vector of dimension 42 (referred as MFCC+f0). Acoustic models are composed of within word triphone HMM models and a 32 components diagonal covariance GMM is used for modeling acoustic emission probabilities. Parameters are shared across different triphones according to a phonetic decision tree. Recognition networks were compiled from trigram language models trained on over one billion words, with a 60K lexicon [8]. Two decoding passes were separated by 3-class MLLR speaker adaptation. Performance of the baseline system on eval06 data is reported in table 2 for both the speaker independent (si) and speaker adapted (sa) models.

Features	CER (si)	CER (sa)
MFCC+f0	27.8	25.8

Table 1: Baseline system performance on eval06 data for 100 hours setup.

In the following, we will consider the use of TANDEM features alone and in concatenation with MFCC+f0 features. The MLP is trained on all the available 100hrs of the training set. The phoneme set is composed of 72 elements. Phoneme posterior probabilities obtained from the MLP are modified according to a Log/KLT transform followed by the reduction of the dimensionality of the TANDEM features to 35 dimensions, and then finally used as conventional features in the HMM/GMM system. The training of MLP has been run at IDIAP while the experiments with the LVCSR system have been done at ICSI.

2.1. Hierarchical Modulation spectrum

The front end used in our previous GALE system is based on the use of MRASTA filtering [9]. MRASTA filtering has been proposed as extension of RASTA filtering trough the use of a two-dimensional band-pass filter. Critical band auditory spectrum is extracted from short time Fourier transform of a signal every 10 ms. Temporal trajectories are filtered with a bank of low-pass filters represented by six first derivatives (G1) and six second derivatives (G2) of Gaussian functions with variance σ varying in the range 8-130 ms (see figure 1 - for details see [9]). The same filters are used for all bands. In the modulation frequency domain, they corresponds to a filter-bank with equally spaced filters on a logarithmic scale thus they provide a multiple-resolution view of the speech dynamics.

Subsequently frequency derivatives are introduced with a context of three-Bark frequency. 19 critical bands are used thus the initial spectro-temporal plane is thus converted into a vector of 432 features. These features form input to a single MLP. MRASTA filtering is particularly effective in case of mismatched acoustic conditions (see [9]).

In a recent work [6] an alternate approach was investigated where, instead of processing all the modulation frequencies by a single MLP, the MRASTA filterbank (six filters) is split into two sets of filterbanks (three filters each): the first set capturing fast modulation frequencies, and the second set capturing slow modulation frequencies as shown in figure 1. The features extracted using these two set of filter banks are then processed in two different ways namely,

1. Parallel processing (Parallel): Two separate MLPs are trained using the fast and the slow modulation frequencies as input. The two MLP outputs are then combined together using a third MLP to estimate a single phoneme posterior estimate.
2. Hierarchical processing (Hier): In hierarchical processing, an MLP is first trained on the fast modulation frequencies. The TANDEM features extracted using this MLP along with the slow modulation frequencies form the input for a second MLP as shown in figure 2

While the parallel processing scheme is insensitive to the order in which the ranges of modulation frequencies are processed, the hierarchical processing scheme depends on the ordering in which ranges of modulation frequencies are introduced. Recognition studies on meeting data showed that hierarchical (sequential) processing can outperform both single classifier and parallel processing.

Meeting recordings are noisy data recorded in a large variety of acoustic environment. In the current work, we investigate whether previous findings are also generalizable to different data (language, condition). In our case, clean condition Mandarin data.

2.1.1. Stand alone TANDEM feature studies

Table 2 reports the performances of TANDEM features obtained training the MLP using the full MRASTA filter-bank (single classifier approach), the fast, the slow ranges of modulation frequencies, parallel processing and hierarchical processing. TANDEM-MRASTA features perform worst then the MFCC+f0 baseline features. The CER of the fast and slow modulation frequency sub-bands is inferior to that of the entire MRASTA filter-bank. On the other hand, when the separate classifiers are combined a consistent improvement is verified: 4.3% (13% relative) in case of parallel combination and 4.6% (14% relative) in case of hierarchical combination. These results are consistent with the findings of the previous study using

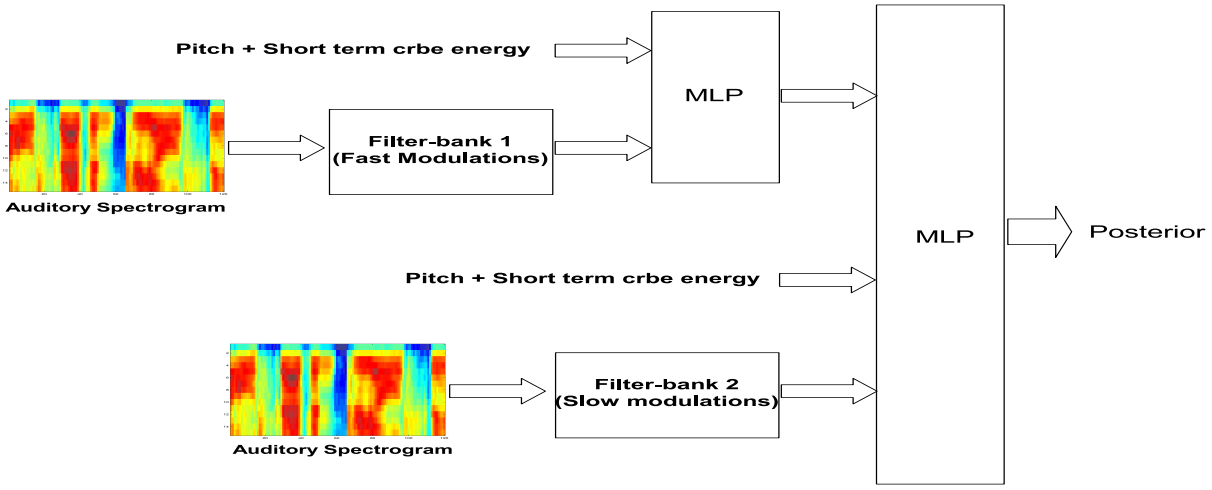

Figure 2: Proposed scheme for the MLP based feature extraction as used in the GALE 2008 evaluation (A-Hier features). The auditory spectrum is filtered with a set of multiple resolution filters that extract fast modulation frequencies. The resulting vector is concatenated with short term critical band energy and pitch estimates and use as input to the first MLP that estimates phoneme posterior distributions. The output of the first MLP is then concatenated with features obtained using slow modulation frequencies, short term critical band energy and pitch estimates and used as input to a second MLP.

Features	CER (si)	CER (sa)
MRASTA	32.4	30.7
Fast	34.5	32.6
Slow	34.9	33.0
Parallel	28.1	26.9
Hier	27.8	26.5

Table 2: CER (expressed in %) for TANDEM features on 100 hours setup.

meeting room data [6]. In other words, although the difference in between the two type of data (meetings recording and broadcast recordings) and the difference in between the Word Error Rate (WER) and the Character Error Rate (CER), the hierarchical processing consistently improves the performance in both cases and yields the best performance.

2.1.2. Concatenation of TANDEM feature with MFCC and pitch

Table 3 reports results of TANDEM features obtined using MRASTA and hierarchical processing in concatenation with MFCC+f0: the resulting system improves drastically w.r.t. the individual feature stream i.e. from 27.8% to 24.4% in case of MRASTA features and to 22.9% in case of hierarchical features. The same conclusions hold also after speaker adaptation thus the improvement obtained with hierarchical features respect to MRASTA is verified after concatenation with MFCC.

Features	CER (si)	CER (sa)
mfcc+f0+MRASTA	24.4	23.1
mfcc+f0+Hier	22.9	21.9

Table 3: CER (expressed in %) for MFCC+f0+TANDEM feature studies on 100 hours setup.

2.2. Hierarchical processing with augmented input features

In this section, we investigate the use of two extra set of features that are added as input to the MLP: the log critical band energy and the smoothed log-pitch estimate.

MRASTA is based on a set of mean-removal filters that extract only the dynamics of the speech signal; this is a suitable property in case of noisy or mismatched data. However in case of clean recordings useful information from the short-term power spectrum is lost. In the following experiments, we append the value of the log critical band energy (19 features per frame) to the MRASTA features. Furthermore we also append the smoothed log-pitch estimate obtained as described in [7]. We refer to this set of features as Augmented MRASTA (A-MRASTA) in case of single classifier approach and Augmented Hierarchy (A-Hier) in case of hierarchical processing. Figure 2 represent the hierarchical processing of the augmented features. Results are reported in table 4. The use of log critical band energy and smoothed log-pitch estimate reduces the CER of approximatively 2% absolute in case of TANDEM features alone and 1% absolute in concatenation with MFCC features. The same conclusions hold also after speaker adaptation.

Features	CER (si)	CER (sa)
A-MRASTA	29.0	26.6
A-Hier	26.4	24.1
mfcc+f0+A-MRASTA	23.4	22.2
mfcc+f0+A-Hier	22.3	21.2

Table 4: CER (expressed in %) for augmented feature studies on 100 hours setup.

3. RWTH evaluation system

In the framework of the GALE project, TANDEM features have been trained on larger amount of data and integrated into the

RWTH speech recognition system. In order to verify the improvements w.r.t. the previous system we report contrastive results on the RWTH GALE 2007 and 2008 evaluation systems trained on 1600 hours of speech. The system consists of two subsystems using different acoustic front-ends. In this paper we will focus on the use of the previosly described MLP features. A more detailed description of the RWTH recognition system could be found in [10, 11].

The significant subsystems are based on MFCC features. These features are normalized by cepstral mean and variance normalization and reduced to 45 dimension by a LDA. This reduced feature vector is augmented with a tonal feature, its first and second derivatives and with MLP posterior features. In training, speaker variations are compensated by speaker adaptive training (SAT) based on CMLLR. The models were enhanced by performing discriminative training with the MPE criterion. The GALE 2007 system used the Hier TANDEM features extracted using MLP trained on 800 hours of speech. The GALE 2008 system used the A-Hier TANDEM features extracted using MLP trained on 1500 hours of speech. Training of MLPs has been run at IDIAP while speech recognition experiments have been run at RWTH.

In recognition, the 3-pass system described in [10] is used. The output of the first recognition is used to estimate the text dependent CMLLR transforms. The second recognition pass is carried out using the CMLLR transformed features and acoustic models trained with SAT. The produced lattices are finally rescored using the full 4-gram language model.

The systems are evaluated on 4 different data sets provided by the GALE project in the last evaluations. The corpora contain up to 2.5h of broadcast news and broadcast conversation. While dev07 is used to tune the system, eval06 and eval07 and dev08 are used for testing. As shown in Table 5, the improved

	CER[%]		
corpus	GALE-2007	GALE-2008	Rel. Improv.
eval06	18.8	15.7	+16%
dev07	12.9	9.6	+25%
eval07	14.1	11.0	+22%
dev08	11.6	9.2	+20%

Table 5: Final recognition results of the subsystem of the GALE-2007 and GALE-2008 evaluation systems, based on MFCC +MLP features. The main improvements is coming from MLP features; other important changes are reported in [10].

features could decrease the character error rate (CER) by approximatively 20% relative for all data set.

4. Summary and Discussion

Table 6 reports the relative improvements w.r.t. the TANDEM-MRASTA baseline both as stand-alone features and in concatenation with MFCC+f0. Experiments refers to the 100 hours system described in section 2. The largest relative improvement is obtained moving from the MRASTA filter-bank to the hierarchical (sequential) processing of the modulation spectrum (14% relative CER reduction with TANDEM features alone and 6% relative in concatenation with MFCC+f0). When critical band energy and pitch are added, the relative improvement becomes respectively 18% and 8%.

The augmented-hierarchical features have been integrated into the RWTH GALE 2008 system and show a significant improvement (20% relative) w.r.t. features used into the RWTH GALE 2007 system.

MLP feature	CER	Rel. Impr. w.r.t. MRASTA
MRASTA	32.4	-
A-MRASTA	29.0	+10%
Hier	27.8	+14%
A-Hier	26.4	+18%
MLP feature+MFCC+f0	**CER**	**Rel. Impr. w.r.t. MRASTA**
MRASTA	24.4	-
A-MRASTA	23.4	+4%
Hier	22.9	+6%
A-Hier	22.3	+8%

Table 6: Relative improvements of the proposed techniques w.r.t. MRASTA features as stand-alone features and in concatenation with MFCC+f0 on eval06 data.

Furthermore the proposed experiments confirm our previous findings i.e. hierarchical (sequential) processing of the modulation spectrum outperforms both the parallel processing and the single classifier approach suggesting that the proposed technique is effective and robust across different tasks and conditions.

5. Acknowledgments

This work was supported by the the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Authors would like to thanks colleagues involved in the GALE project at IDIAP, ICSI, RWTH and SRI especially Arlo Faria and Andreas Stolcke.

6. References

[1] Hermansky H., Ellis D., and Sharma S., "Connectionist feature extraction for conventional HMM systems.," *Proceedings of ICASSP*, 2000.

[2] Bourlard H. and Morgan N, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Publishers, 1994.

[3] Morgan N. et al., "TRAPping conversational speech: Extending TRAP/Tandem approaches to co nversational telephone speech recognition.," *Proceedings of ICASSP 2004*.

[4] Hermansky H., "Should recognizers have ears?," *Speech Communications*, vol. 25, pp. 3–27, 1998.

[5] Kingsbury B.E.D., Morgan N., and Greenberg S., "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, pp. 117–132, 1998.

[6] Valente F. and Hermansky H., "Hierarchical and parallel processing of modulation spectrum for asr applications," *Proceedings of ICASSP 2008*.

[7] Lei X. et al., "Improved Tone Modeling for Mandarin Broadcast News Speech Recognition.," *Proceedings of Interspeech, 2006*.

[8] Hwang M.-Y., Peng G., Wang W., Faria A., and Heidel A., "Building a Highly Accurate Mandarin Speech Recognizer," *Proceedings ASRU, 2007*.

[9] Hermansky H. and Fousek P., "Multi-resolution RASTA filtering for TANDEM-based ASR.," in *Proceedings of Interspeech 2005*, 2005.

[10] Plahl C. et al., "Recent improvements of the RWTH GALE Mandarin LVCSR system.," *Proc. Interspeech 2008*.

[11] Plahl C. et al., "Development of the GALE 2008 Mandarin LVCSR System," *Submitted to Interspeech 2009*.