FEATURE EXTRACTION USING NON-LINEAR TRANSFORMATION FOR ROBUST SPEECH RECOGNITION ON THE AURORA DATABASE

Sangita Sharma[1,*], *Dan Ellis*[2], *Sachin Kajarekar*[1], *Pratibha Jain*[1] *and Hynek Hermansky*[1,2]

[1]Oregon Graduate Institute of Science and Technology, Portland, Oregon , USA
[2]International Computer Science Institute, Berkeley, California, USA.

ABSTRACT

We evaluate the performance of several feature sets on the AURORA task as defined by ETSI. We show that after a non-linear transformation, a number of features can be effectively used in a HMM-based recognition system. The non-linear transformation is computed using a neural network which is discriminatively trained on the phonetically labeled (forcibly aligned) training data. A combination of the non-linearly transformed PLP, MSG and TRAP features yields a 63% improvement in error rate as compared to a baseline MFCC features. The use of the non-linearly transformed RASTA-like features, with system parameters scaled down to take into account the ETSI imposed memory and latency constraints, still yields a 40% improvement in error rate.

1. AURORA TASK

The AURORA task [12] has been defined by the European Telecommunications Standards Institute (ETSI) as a cellular industry initiative to standardize a robust feature extraction technique for a distributed speech recognition framework. The initial ETSI task uses the TI-DIGITS database downsampled from the original sampling rate of 20kHz to 8 kHz and normalized to the same amplitude level. Four different noises - exhibition hall noise, babble noise, suburban train noise and moving car noise have been artificially added to different portions of the database at signal-to-noise (SNR) ratios ranging from clean, 20dB to 0dB in decreasing steps of 5dB.

The training set consists of 8440 different utterances split equally into 20 subsets of 422 utterances each. Each split has one of the four noises added at one of the five SNRs (clean, 20dB, 15dB, 10dB and 5dB). The test set consists of 4000 test files divided into four sets of 1000 files each. Each set is corrupted with one of the four noises at 6 SNR levels (clean, 20dB, 15dB, 10dB, 5dB and 0dB), resulting in a total of (4 x 1000 x 6) 24,000 test utterances.

The recognition system for this evaluation has been fixed to be a toolkit-based (HTK) HMM system with eleven whole word digit models, each comprising of 16 states with 3 mixtures per state. Two silence models, one with 3 states and 3 mixtures to model the utterance beginning and end silence, and the other with 1 state and 6 mixtures to model the interword silence have also been used.

In spite of some drawbacks of the current AURORA task such as the matched test and training conditions, or the absence of natural level variations and variable linear distortions, the AURORA task is of interest since it can demonstrate the potential benefits of using noise-robust feature extraction techniques towards improving the recognition performance on a task which (though with matched training and test conditions) has substantial variability due to different types of additive noise at several SNRs.

*Now with Intel Corporation.

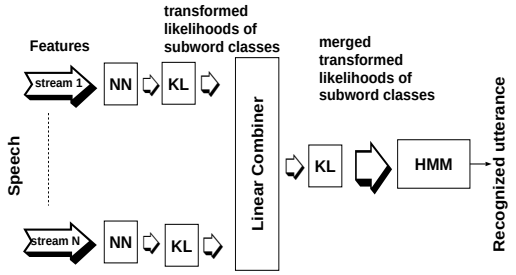

Figure 1: Feature Extraction block diagram.

2. TEMPORAL-BASED FEATURES

Spectral-based features such as mel-frequency cepstral coefficients (MFCC), perceptual linear predictive coefficients (PLP) etc., form the basis of most feature extraction techniques in automatic speech recognition (ASR) systems. These features characterize the spectral envelope in a short-time frame (typically 10ms) of speech. A drawback of the spectral features is that they are quite sensitive to changes in the communication environment such as changes in communication channels or environmental noise. Subsequently, the performance of recognizers based on spectral features rapidly degrades in realistic communication environments.

Psychoacoustic studies (reviewed in [4]) suggest that the peripheral auditory system in humans integrates information from much larger time spans than the temporal duration of the frame used in speech analysis. This time span is of the order of several hundred milliseconds (around 200ms). Several emergent noise robust techniques (reviewed in [5]) now employ short-term feature vectors which integrate information from such medium-time spans. In this paper, we evaluate three such recently proposed techniques.

2.1. Temporal LDA based RASTA-like features

This technique introduced in [14] employs a data-driven approach to the design of RASTA-like [6] filters on the time trajectories of log critical-band energies. The linear discriminant analysis (LDA) technique is used to optimize the linear discriminability between the phoneme classes in the presence of undesirable within-class variability (such as those introduced by speaker, channel, context and environmental noise). The vector space for LDA is constructed from around 1 sec segments (101 points at 10ms frame rate) of the time trajectory of a single log critical-band energy, each of which is labeled with respect to the center frame. These filters are derived from around 3 hours of phonetically hand-labeled OGI-Stories database [2]. Since, the AURORA task includes variability due to environmental noise, volvo-car noise from the NOISEX-92

database [15], at a signal-to-noise ratio of 10dB, was artificially added to the stories database. The filters are derived from the 15 Bark-spaced critical band energies [3]. The first three principal vectors obtained from the LDA analysis for each of the 15 critical bands are used as 101-tap FIR RASTA-like filters. The use of the three RASTA-like filters per critical band results in a 45-dimensional feature vector per frame. Karhunen-Loeve (KL) transform derived on the training set of the AURORA database is then used to reduce the dimensionality of the feature vector to 39 parameters (retaining 99.9% of the variability) and to diagonalize the feature vector.

2.2. Modulation-Filtered Spectrogram (MSG) features

The modulation-filtered spectrogram is a robust speech representation for ASR [10]. The robustness of the representation is based on two signal-processing strategies: 1) the emphasis of changes in the spectral structure of the speech signal (measured with critical-band-like resolution) occurring at rates of 16 Hz or less, 2) the adaptation to slowly-varying components of the speech signal that functions as a form of automatic gain control (AGC). The particular form of the algorithm used in these experiments uses two modulation filter banks, covering roughly the 0-8 Hz and 8-16 Hz modulation bands. The modulation-filtered spectrogram features were computed as described in [10]. The MSG feature extraction technique is carried out on 14 Bark-spaced critical band energies and yields a 28-dimensional feature vector per frame.

2.3. TempoRAl Pattern (TRAP) features

The temporal pattern (TRAP) feature extraction technique is a technique to extract temporal information from the speech signal to improve the noise robustness of ASR systems [8]. This technique uses two concepts hypothesized to occur in human hearing: 1) independent processing at individual frequency channels, and 2) temporal processing over medium time (syllable-length) spans. In each of the 15 Bark-spaced critical bands, a one second (101 point) long temporal vector of logarithmic energies is used as input to a neural network (multi-layer perceptron (MLP)) for estimating the probability of the phoneme at the center of this vector. The MLPs in each critical band are trained on task-independent phonetically labeled OGI-Stories database. The phonetic probability estimates thus obtained independently from all the critical bands are further non-linearly merged using a MLP which is trained on the task-specific training data. This vector of merged probability estimates comprises the TRAP feature set. The dimensionality of this vector depends on the number of phonemes present in the database.

3. NON-LINEAR TRANSFORMATION OF FEATURES

In [9] it was shown that a MLP trained on the task-specific training data can be used to derive a mapping from any feature set to the logarithmic likelihoods of context-independent phonemes. This mapping is obtained by discriminatively training the MLP on the phonetic labeled training data using a softmax activation function on the output layer. The outputs of the MLP represent estimates of the phoneme posterior probabilities [11]. In a well-trained MLP, the right class typically has an estimate close to 1, while the other class estimates are close to zero. This results in a highly non-Gaussian distribution of the parameters in the feature vector. Such a non-Gaussian distribution can violate the mixture-of-Gaussian assumption typically used in a HMM system. One way to make the distribution closer to Gaussian is to remove the output softmax non-linearity from the trained net. The parameters of such linearized outputs are then further diagonalized through Karhunen-Loeve (KL) transform for subsequent HMM modeling using diagonal covariance matrices.

The above described non-linear transformation requires that the training data be phonetically labeled. For the AURORA database, we obtained the initial phonetic segmentation for the training part by forced alignment using a hybrid (HMM/MLP) recognition system trained on the OGI Numbers task [2]. The Numbers database consists of the same eleven digit vocabulary as the AURORA database. This initial segmentation was further improved using embedded training [11] on the AURORA database. The AURORA database was labeled in terms of 24 monophone classes that describe the digits.

4. EXPERIMENTS

4.1. Combining multiple features

The features used in all the experiments described in this paper have been derived from speech frames of length 25ms with a frame rate of 10ms. Each frame has been windowed using a Hamming window function.

4.1.1. Baseline system

The baseline system as defined by ETSI uses as features, 13 MFCC coefficients along with their delta and acceleration coefficients. The MFCC coefficients are derived from 23 mel-spaced triangular filters. The baseline system thus uses a 39-dimensional feature vector per frame. Table 1 shows the performance of the baseline system at the different SNRs averaged over the 4 noises.

4.1.2. PLP-based system

The PLP feature extraction technique [3] differs from the MFCC feature extraction technique mainly in the use of 15 Bark-spaced critical bands followed by cepstral coefficient computation from the autoregressive modeling of the critical band power spectrum. Table 1 shows that the performance of the PLP features (39-dimensional vector consisting of 13 cepstral coefficients along with delta and acceleration coefficients) is close to that of the baseline system.

Table 1 also shows the performance of the non-linearly transformed PLP features. The per frame input to the MLP for this non-linear transformation consisted of 9 frames (current frame, 4 frames in the past and 4 frames in the future) as commonly used in a hybrid ASR system [1] — i.e. (39 x 9) 351 inputs, 480 hidden units and 24 outputs. PLP-NN refers to the non-linearly transformed PLP features. It is the seen that these give around 48% reduction in error as compared to the PLP features and the baseline features.

4.1.3. Temporal LDA base RASTA-like features

In Table 1, LDA refers to the performance of the 39 RASTA-like features when used directly for recognition, while LDA-NN refers to the performance of these features after the non-linear transformation into a 24-dimensional feature vector. Similar to

Table 1: Word error rate (%) for various features and combinations.

FEATURES	Clean	20dB	15dB	10dB	5dB	0dB	Average reduction in error (20-0dB)
Baseline	1.5	2.7	3.8	7.3	16.8	41.6	
PLP	1.2	2.7	4.1	7.5	16.8	40.9	-2
PLP-NN	1	1.4	2.1	3.7	8.4	22.4	48
LDA	1.5	2.4	3.8	6.7	13.9	28.3	14
LDA-NN	1.1	1.5	2.1	3.5	8.1	19.9	49
MSG	6.0	5.7	7.8	12.0	23.2	42.9	-64
MSG-NN	1.0	1.3	1.8	3.5	8.5	23.5	50
TRAPs	3.4	2.6	3.1	5.3	10.8	27.3	24
PLP-NN + LDA-NN	0.9	1.2	1.9	3.4	7.9	20.2	53
PLP-NN + MSG-NN	0.6	1.0	1.4	2.8	7.0	19.9	60
PLP-NN + TRAP	0.9	1.1	1.7	2.9	7.2	18.5	57
LDA-NN + TRAP	1.1	1.2	1.8	3.2	7.7	19.2	54
PLP-NN + LDA-NN + MSG-NN	0.8	1.0	1.5	2.8	7.2	19.0	59
PLP-NN + MSG-NN + TRAP	0.7	0.9	1.3	2.7	6.5	17.5	63
PLP-NN + LDA-NN + TRAP	0.8	1.0	1.5	2.9	6.9	18.1	60
PLP-NN + LDA-NN + MSG-NN + TRAP	0.8	1.0	1.4	2.6	6.6	18	62

Table 2: Word error rate (%) for the temporal-LDA system and reduced-complexity variants.

FEATURES	Size of MLP	Clean	20dB	15dB	10dB	5dB	0dB	Average reduction in error (20-0dB)
LDA101	(39x9:500:24)	1.1	1.5	2.1	3.5	8.1	19.9	49
LDA41	(39x9:500:24)	1.1	1.6	2.3	3.9	8.3	19.6	46
LDA41-Q-ds3	(39x3:200:24)	1.2	1.6	2.4	4.5	9.9	24.8	40

the PLP-NN system the per-frame input to the MLP consists of 9 frames (351 inputs), 500 hidden units and 24 outputs. It is seen that the LDA system results in around 14% reduction in error as compared to the baseline system, while the LDA-NN system results in 49% reduction in error.

4.1.4. MSG features

As seen from Table 1, the 28-dimensional MSG features when used directly for recognition give worse performance than the baseline system (64% increase in error). However, the non-linearly transformed features, MSG-NN, (using an MLP with (28 x 9frames) 252 inputs, 480 hidden units and 24 outputs) yield a 50% reduction in error as compared to the baseline system.

4.1.5. TRAP features

As used in [8], each of the critical band MLPs in the TRAP system used a 101 point input, 300 hidden units and 24 outputs. The combiner MLP uses (24 x 15) = 360 inputs, 300 hidden units and 24 output units. As seen from Table 1, these TRAP features yield a 24% reduction in error as compared to the baseline system.

4.1.6. Feature combination

If two or more systems yield complementary information an effective combination of the outputs of these systems can yield performance better than any single system [13] (see Fig. 1) The spectral (PLP-based) system and each of the temporal-based systems generally make considerable number of complementary errors [5]. Hence we tested several combinations of the above four features as shown in Table 1. The combination was done by averaging the corresponding non-linearly transformed features sets prior to orthogonalization. It is seen that each of the combinations gives an improvement in performance as compared to any single system. The best system which uses a combination of the non-linear transformed PLP and MSG features along with the TRAP features gives the best (around 63%) reduction in error as compared to the baseline system.

4.2. RASTA-like features with scaled down non-linear transformation

The above system assumed no constraints in terms of computation, available memory and latency. The time-delay introduced by the above features was of the order of around 500ms. However, the Aurora task had the constraints that the latency should not exceed 250ms and the ROM requirement should be around 15 kwords (30KB). It is possible for the above system to meet these requirements, if we allow for some stages of the feature-extraction to be done on the server-end of the DSR system. In that case only the critical band analysis, which is used in all the above techniques, can be done on the front-end, with the non-linear transformation moved to the server-end.

However, to evaluate the effect of scaling down the entire feature extraction to meet the ETSI requirements, we chose to use our system based only on the non-linearly transformed RASTA-like features (LDA). The scaling down was done as follows:

- To reduce the latency to 200ms, the 101-tap filters were approximated by 41-tap filters. A least square approximation of the frequency response of the 101-tap filters by 41-tap filters was used.
- To reduce memory requirements, the 45 FIR filters (3 filters per critical band) were replaced by 3 FIR filters (same filters irrespective of the critical bands) by taking the mean of the

impulse response of each set of 15 filters. This approximation is possible since the different critical bands have similar filters [14].

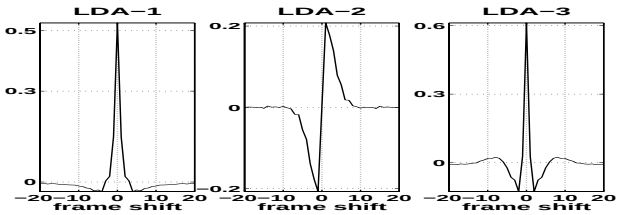

Figure 2: Impulse responses of temporal LDA filters.

- The 3 FIR-filters show a band-pass frequency characteristic with upper cut-off frequency around 12-16 Hz. This indicates that the critical-band time-trajectories can be downsampled by a factor of 3 (original sampling frequency is 100Hz at 10ms frame-rate) [7]. In other words the time trajectories are filtered for every third frame only, thus reducing computation.
- The basis of the KL transform are quantized so as to be accurate to the first decimal value. With such a quantization, each of the (45 x 39) values in the KL transformation can be stored in a single byte.
- The 9 frame input to the non-linear transformation MLP is reduced to a 3 frame input (current frame, third frame in the past and future which are actually consecutive frames after downsampling). Further we reduce the size of the hidden layer to 200 units. This scaling down results in a transformation matrix of (39x3x200 + 200x24 + 200 + 24) 28,424 parameters as compared to the original size of (39x9x500 + 500x24 + 500 + 24) 188,024 parameters. Thus we obtain a reduction by a factor of 7 in the number of parameters, each of which can be further quantized so as be represented in a single byte. The additional latency introduced by the 3 frame input is 30ms (total of 230ms)
- The downsampled features are then interpolated before recognition [7].

From the Table 2 it is seen that the use of 3, 41-tap filters (LDA41) results in only slight degradation in performance as compared to the 45, 101-tap filters (LDA101 same as LDA-NN in Table 1). Further scaling down this system as described above, still yields in a 40% reduction in error as compared to the baseline system.

5. CONCLUSION

Our results show that non-linear transformation of features can significantly improve the ASR system performance. Further, combining different non-linearly transformed feature sets, especially complementary feature sets such as those based on spectral and temporal processing, yields a further improvement in performance. Finally, we show that it is possible to considerably scale down our systems to meet practical constraints, while still maintaining noticeable robustness in recognition performance.

6. ACKNOWLEDGMENT

The research was supported by DoD under MDA904-98-1-0521, NSF under IRI-9712579, by an industrial grant from Intel Corporation, and by European Union ESPRIT LTR project Respite (28149).

7. REFERENCES

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

[2] R. A. Cole, M. Noel, T. Lander, and T.Durham. New telephone speech corpora at CSLU. *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'95)*, 1:821–824, September 1995.

[3] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[4] H. Hermansky. Exploring temporal domain for robustness in speech recognition, invited paper. *Proceedings of the 15th International Congress on Acoustics*, 3:61–64, 1995.

[5] H. Hermansky. Should recognizers have ears? *Speech Communication, Invited paper*, 25(1-3):3–27, 1998.

[6] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, 1994.

[7] Hynek Hermansky and Pratibha Jain. Down-sampling speech representation in ASR. *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'99)*, Sep 1999.

[8] Hynek Hermansky and Sangita Sharma. TempoRAl Patterns (TRAPS) in ASR of noisy speech. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'99)*, 1:289–292, March 1999.

[9] Hynek Hermansky, Sangita Sharma, and Pratibha Jain. Data-derived nonlinear mapping for feature extraction in HMM. *Proceedings of the Workshop on Automatic Speech Recognition and Understanding, to be published*, Dec 1999.

[10] Brian E. D. Kingsbury. Perceptually-inspired signal processing strategies for robust speech recognition in reverberant environments. *Ph.D. Thesis, University of California, Berkeley, CA*, 1999.

[11] N. Morgan and H. Bourlard. An introduction to hybrid HMM/connectionist continuous speech recognition. *IEEE Signal Processing Magazine*, 12(3):25–42, May 1995.

[12] David Pearce. Aurora project: Experimental framework for the performance evaluation of distributed sppeech recognition from-ends. Sep 1998.

[13] Sangita Sharma, Pieter Vermeulen, and Hynek Hermansky. Combining information from multiple classifiers for speaker verification. *Proceedings of the Speaker Recognition and its Commercial and Forensic Applications Workshop (RLA2C)*, pages 115–119, April 1998.

[14] S. van Vuuren and H. Hermansky. Data-driven design of RASTA-like filters. *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'97)*, pages 409–412, 1997.

[15] Andrew Varga and Herman J.M. Steeneken. Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251, 1993.