BELL LABS APPROACH TO AURORA EVALUATION ON CONNECTED DIGIT RECOGNITION

Jingdong Chen Dimitris Dimitriadis Hui Jiang Qi Li
Tor André Myrvoll Olivier Siohan Frank K. Soong

Multimedia Communications Research Laboratory
Bell Labs, Lucent Technologies
Murray Hill, NJ 07974, USA

ABSTRACT

In this paper we study various front-end features, modeling and adaptation algorithms on the Aurora 3 databases, including auditory, moment, and AM-FM modulation features, context-dependent digit models, segmental K-means training, discriminative training, and model adaptations. The evaluation results on Aurora 3 are presented with a brief summary of our Aurora 2 results.

1. INTRODUCTION

The Aurora evaluation is for researchers to test their algorithms on noise robustness and compare results measured on the same databases. So far, there are two tasks on the Aurora evaluation, Aurora 2 and 3, both are for connected digit recognition. While the Aurora 2 databases use the controlled experiments by adding noise digitally to clean English digit strings [1], the Aurora 3 databases are collected in a real-world car environment in 4 languages. In this paper, we report our evaluation results on two of the languages, Spanish and German.

2. BELL LABS APPROACHES

In this section, we present our baseline system then describe the different feature sets that have been used for this evaluation. Alternative training strategies and acoustic model adaptation techniques are also reviewed.

A. Context-Dependent Model: Similar to last year approach [1], we have decided to use context-dependent (CD) digit models, together with Bell Labs recognition engine as backend. This contrasts with the official Aurora backend that is based on whole-word digit models and the HTK engine. The official backend setup typically leads to poorer results, especially in larger databases, and we believe that a better baseline is beneficial to properly study the effect of different front-ends on the final recognition performance.

Last year, we investigated several approaches to build CD digit models. Given the limited amount of training data, especially in the Aurora3 databases, it is required to rely on some tying techniques to build CD digit models. The Head-Body-Tail digit model structure (HBT) assumes that CD digit models are built by concatenating a left-context-dependent unit (head) with a context-independent unit (body) followed by a right-context-dependent unit (tail). For example, assuming that the lexicon contains 10 digits plus a silence model, each digit model consists of a set of 1 body, 11 heads and 11 tails (representing all left/right contexts) [2]. We typically model each head and tail with a 3-state HMM, while a 4-state HMM is used for each body. Most of the experiments done this year have been based on the HBT structure. CD digit models can also be built as triphone models using a decision tree. This is the approach we introduced last year [1], and some of this year experiments have been carried out using this model topology.

B. Auditory Feature: The new auditory front-end in our recognition system was developed to mimic the robust human hearing in adverse acoustic environments [3, 4]. In the front-end, efficient signal processing functions were implemented to satisfy both real-time and computation cost requirements. Based on the analysis of the outer and middle ear, a transfer function was constructed to replace the commonly used preemphasis filter, and then a new set of digital auditory filters, which simulate auditory filtering in the cochlea, replaces those used in the MFCC and PLP. The auditory feature extraction procedure consists of: an outer-middle-ear transfer function, FFT, frequency conversion from linear to the Bark scale, auditory filtering, nonlinearity, and discrete cosine transform (DCT). In our previous study[3], the feature has been evaluated in two tasks: connected-digit and large vocabulary, continuous speech recognition under various noise conditions, using both handset and hands-free data in landline and wireless transmission with additive car and babble noise. Compared with the LPCC, MFCC, MEL-LPCC, and PLP features, the auditory feature achieved significant performance improvement

in the connected-digit and the Wall Street Journal tasks[3]. The major improvement is due to the new auditory filters.

C. Normalized Moment Feature: Spectral subband centroid (SSC) feature was first proposed for speech recognition [5] and then extended by incorporating the dynamic SSC [6]. Inspired by the concept of SSC, a more generalized, normalized subband spectral moments (NSSM) was proposed for noisy speech recognition [7]. In brief, we divide the full band into total of I subbands, the NSSM of order *p* for the *i*-th subband is defined by

$$NM^p(t,i) = \frac{M^p(t,i)}{M^0(t,i)} = \frac{\int_0^\pi \omega^p w_i(\omega) P(t,\omega) d\omega}{\int_0^\pi \omega^p P(t,\omega) d\omega} \quad (1)$$

where $M^p(t,i)$ indicates the *i*-th subband moment of order p at time t, $w_i(\omega)$ is the frequency response of the *i*-th bandpass filter, and $P(t,\omega)$ the short-term power spectrum of the speech signal at frame t. In this paper, the NSSMs of order 2 are applied to the Aurora 3 task with a HBT model structure. Each frame is represented in the energy and 12 NSSM features estimated from unsmoothed FFT power spectrum using 12 rectangular band-pass filters with 50% overlap, distributed along a linear scale [7].

D. AM-FM Modulation Feature: There exists some evidence that leads us to analyze speech signals in their amplitude and frequency modulation(AM-FM) components which vary continuously. We estimate these variations using the Teager-Kaiser energy-tracking operator [9] energy separation algorithm (ESA). This algorithm demodulates narrowband AM-FM signals by tracking the physical energy implicit in the source and separating it into amplitude and frequency components. For implementation purpose, we first convert discrete samples into corresponding continuous signals by using smoothing splines [8] before we apply the continuous-time ESA. In speech recognition experiments we create a hybrid feature vector by augmenting the standard auditory feature vector with the FM-AM information.

The extraction of modulation features is as follows: (i) A filterbank of six, 50% overlapped Gabor bandpass filters with center frequencies near the averaged formant frequencies, is used. (ii) The output from each Gabor bandpass filter is demodulated via the Spline-ESA into instantaneous amplitude, $a(t)$, and frequency, $f(t)$, components. (iii) The lowpassed $a(t)$ and $f(t)$ are segmented into 30 ms frames, updated every 10 ms. (iv) For each analysis frame and each band, the weighted mean F_w and standard deviation B_w of the instantaneous frequency signal are estimated:

$$F_w \equiv \frac{\int_1^T f(t)a^2(t)dt}{\int_1^T a^2(t)dt} \quad (2)$$

$$B_w^2 \equiv \frac{\int_1^T [(\dot{a}(t)/2\pi)^2 + (f(t) - F_w)^2 a^2(t)]dt}{\int_1^T a^2(t)dt} \quad (3)$$

where T is frame length. (v) The FM percentage in each band, $K_i = B_w/F_w$, is computed.

E. Channel-Dependent (ChD) Model: In Aurora 3, the well-matched (WM) training scenario pools together heterogeneous data, close-talking and hands-free microphone recordings, based on a multi-style training paradigm. Since the well matched scenario is based on two very distinct environments, close-talking and hands-free. We build condition-dependent models by adapting a condition independent model. The condition independent model is built on the official WM training set. This model can then be adapted to both the hands-free and close-talking data by using SMAP [13].

Since we now have 2 models available for recognition, we need to decide which one should be used for each test utterance. This can be simply done by running recognition in parallel using the hands-free and close-talking models, and select the recognition hypothesis with the highest score. Using MFCC feature and HBT digit models, the baseline word error rate (WER) on the Spanish data is 4.0% on the WM test set. When using an oracle to select which model should be used for each test utterance, we obtain 0.6% WER on the close-talking data, and 6.5% WER on the hands-free data, for an average of 3.5% WER. By selecting the model based on the highest recognition score, the average WER on WM is 3.5%, similar to the oracle result. This illustrates the heterogeneity of the data, and indicates that the channel can be identified with high accuracy on this database.

F. MCE/GPD Training: Minimum classification error (MCE) is a discriminative training objective that associates with classification error directly [11]. The corresponding training algorithm named generalized probabilistic descent (GPD) algorithm has been used in Bell Labs for many years. Our large database evaluations have indicated that the MCE/GPD training can reduce training errors significantly, especially for the connected digit recognition tasks [3]. The algorithm has been applied to this Aurora 3 evaluation.

G. Adaptation: There are two conditions that make unsupervised adaptation for this task particularly challenging. For one, no information of speaker's identity or noise conditions is available. This means that a new, adapted model has to be estimated for every utterance. Second, the amount of adaptation data can be very scarce – as short as a monosyllable digit token.

The above conditions call for a transformation based approach like MLLR. The reason for this is that approaches like MAP [14] adaptation tend to adapt the erroneously recognized models away from the correct direction and the unseen(i.e., unrecognized) digits are left unmodified. But even the use of a single, global MLLR transformation can still be problematic when data for adaptation is in the order of a few hundred milliseconds.

In this work we use the transformation-based approach

described in [15]. Here the mean vectors of mixture components, μ, are adapted using a diagonal transformation matrix, D, plus a translation vector, b, using the relation $\hat{\mu} = D\mu + b$. The transformation $\{D, b\}$ is estimated in a maximum a posteriori sense, where the prior distribution, $g(D, b|\Phi)$, is based on a hybrid of hierarchical and empirical Bayes approaches. As complexity of the adaptation mapping can be controlled smoothly using the transformation prior distributions, this approach scales well for adaptation data ranging from mono-syllable digits to digit strings.

3. AURORA 3 EVALUATION

We evaluated the above techniques on the Aurora 3 Spanish and German databases. All the features in our evaluation are 39-dimensional vectors including 12 cepstral coefficients, energy (c_0 of DCT or short-term energy), plus their first and second order time derivatives, with the exception that the AM-FM feature uses 6 extra feature in additional to the auditory cepstral coefficients plus their derivatives. For a fair comparison, all the features were evaluated using the same HBT model structure. Although we did not use the standard HTK backend, the fixed model structure with the same model topology and the same number of parameters serves the same purpose for evaluation.

To speedup the training process, all the training data were first segmented into the HBT unit level by pre-trained HBT models on the LPCC feature. Initial models were trained based on the initial segmentation using the segmental K-means algorithm [10] and the auditory feature. The trained models were then used to segment the training data again. The final models for different features were trained based on the second segmentations. Compared to the forward-backward algorithm, the segmental K-means is much faster. Each approach is evaluated under three training and testing conditions as defined by the Aurora 3 task as well-matched (WM), medium-matched (MM), and high-mismatched (HM) experiments.

Spanish Database: Several features were first evaluated on the Spanish database. As shown in Table 1, the approaches included: linear predictive coding coefficients (LPCC), mel-frequency cepstral coefficients (MFCC), normalized moment (Moment), auditory (Auditory or Aud.), and combined auditory and AM-FM modulation (Aud. + AM-FM) features. The training algorithm is maximum likelihood estimate (MLE). The HBT models were used except the MFCC case, where CD model was used with noise compensation on the HM condition [18]. The average is weighted average as defined in the standard spreadsheet[1].

Furthermore, the generalized probabilistic descent (GPD) algorithm was applied to the LPCC, MFCC, and Auditory features to train the MCE models. The online adaptation algorithm described above was applied to the MFCC and Auditory features to further improve the performances. The results are listed in Table 2.

[1] Weighted Average = 0.40WM + 0.35MM + 0.25HM.

Table 1: Comparisons on Different Features (%)

Spanish	WM	MM	HM	Ave.[1]
Aurora Baseline	86.9	73.7	42.2	71.11
LPCC + MLE	94.5	89.0	80.9	89.18
MFCC + MLE	96.2	91.2	85.5	91.77
Centroid + MLE	94.1	89.0	82.7	89.47
(Aud + AM-FM)+MLE	95.2	88.3	85.2	90.29
Aud + MLE	96.0	91.0	86.2	91.80

Table 2: Comparisons on Training and Adaptation (%)

Spanish	WM	MM	HM	Ave.[1]
LPCC + MLE + GPD	95.3	89.7	83.4	90.37
MFCC + MLE + GPD	96.2	89.3	83.3	90.56
MFCC + GPD + Adapt.	96.3	89.9	85.0	91.24
MFCC + ChD/MLE	96.5	–	–	–
Aud + MLE + GPD	96.2	91.0	87.1	92.11
Aud + GPD + Adapt.	96.2	91.1	87.7	92.14
Aud + ChD/MLE	95.6	–	–	–

German Database: The procedure of the German database evaluation is similar to the Spanish one and the results are listed in Tables 3 and 4. The training and testing utterances in the German database are much less than the Spanish one. Due to the small amount of data, the evaluation results may not be as significant as using a larger database.

Table 3: Comparisons on Different Features (%)

German	WM	MM	HM	Average[1]
Aurora Baseline	90.6	79.1	74.3	82.50
LPCC + MLE	91.5	80.8	84.5	86.01
MFCC + MLE	92.2	80.3	85.8	86.44
Moment + MLE	90.3	77.5	83.6	84.15
Aud + MLE	92.9	83.5	86.6	88.04

4. SUMMARY OF EVALUATIONS

Aurora 2 Evaluation: Our Aurora 2 system differs radically from last year system [1]. We followed an approach fairly similar to Ellis' work [17]. The basic idea is to use a neural network to derive the posterior probability of the head, body and tail portion of a digit segment. The posterior probability vector is then used as feature vector and an HBT recognizer is built. We built two systems based on this principle. In the first one, the input of the neural network consists of a window of consecutive MFCC vectors, while the output consists of nodes representing the head,

Table 4: Comparisons on Training and Adaptation (%)

German	WM	MM	HM	Ave.[1]
LPCC + MLE + GPD	92.3	80.7	85.0	86.42
MFCC + MLE + GPD	92.6	81.0	85.5	86.77
MFCC + GPD + Adapt.	92.6	81.5	86.4	87.17
MFCC + ChD/MLE	92.5	–	–	–
Aud + MLE + GPD	92.9	83.5	86.6	88.04
Aud + GPD + Adapt.	92.9	83.5	86.6	88.04
Aud + ChD/MLE	92.3	–	–	–

Table 5: Summary of Aurora 2 Average Word Accuracy (%)

Training Mode	Set A	Set B	Set C	Overall
Multicondition	93.72	92.84	93.34	93.29
Clean Only	86.95	87.71	83.98	86.66
Average	90.33	90.27	88.66	89.98

body and tail of all digits. The second system is similar to the first one, except that the delta MFCCs are also fed to the neural network. We also built a baseline HBT system in a standard way using MFCC features, as described in [1]. In these 3 systems, and for the clean only training condition, the input test MFCC vectors have been processed using the noise compensation technique described in [18]. The recognition results of these 3 systems are finally combined using ROVER, and are presented in Table 5. This illustrates that the use of the non-linear feature extraction significantly improves the recognition performance. Reader are refered to [16] for full details.

Aurora 3 Evaluation: The summaries of Aurora 3 word error rates and relative improvement are listed in Tables 6 and 7 followed the standard spreadsheet. The numbers for the Spanish and German columns are from Tables 2 and 4, respectively.

Table 6: Summary of Aurora 3 Word Error Rate (%)

	Spanish	German
Well (x40%)	3.5	7.1
Mid (x35%)	8.8	16.5
High (x25%)	12.3	13.4
Overall	7.56	11.97

Table 7: Summary of Aurora 3 Relative Improvement (%)

	Spanish	German
Well (x40%)	50.42	19.32
Mid (x35%)	47.27	12.97
High (x25%)	74.61	50.06
Overall	55.37	24.78

5. REFERENCES

[1] M. Afify, H. Jiang, F. Korkmazskiy, C.-H. Lee, Q. Li, O. Siohan, F. K. Soong, and A. C. Surendran, "Evaluating the aurora connected digit recognition task – A Bell Labs approach," in *Proc. of EuroSpeech*, pp. 633–637, Sept. 2001.

[2] W. Chou, C.-H. Lee, and B.-H. Juang, "Minimum Error Rate Training of Inter-word Context Dependent Acoustic Model Units in Speech Recognition," *Proc. ICSLP-94*, Sept. 1994.

[3] Q. Li, F. K. Soong, and O. Siohan, "An Auditory System-Based Feature for Robust Speech Recognition," *Proce. of Eurospeech*, vol. 1, pp. 619–622, 2001.

[4] Q. Li, F. K. Soong, and O. Siohan, "A High-Performance Auditory Feature for Robust Speech Recognition," *Proc. ICSLP*, 2000.

[5] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *Proc. IEEE ICASSP* 1998, vol. II, pp. 617-620.

[6] J. Chen, Y. Huang, Q. Li, and K. K. Paliwal, "Dynamic spectral subband centroid features for robust speech recognition," submitted to the *IEEE Signal Proc. Letters*.

[7] J. Chen, Y. Huang, Q. Li, and F. K. Soong, "Recognition of noisy speech using normalized moments," submitted to ICSLP'2002.

[8] D. Dimitriadis and P. Maragos, "An Improved Energy Demodulation Algorithm Using Splines", *Proc. ICASSP-01*, May 2001.

[9] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy seprattion in signal modulations with application to speech analysis", *IEEE Trans. Signal Proc.* vol. 41, pp. 3024-3051, Oct. 1993.

[10] L. R. Rabiner, J. G. Wilpon and B.-H. Juang, "A segmental k-means training procedure for connected word recognition", *AT&T Technical Journal*, vol. 65, no. 3, May/June 1986.

[11] B.-H. Juang, W. Chou, and C.-H. Lee. "Minimum classification error rate methods for speech recognition," *IEEE Trans. on Speech and Audio Process.*, vol. 5, no. 5, pp. 257-265, May 1997.

[12] W. Reichl and W. Chou, "Robust decision tree state tying for continuous speech recognition," *IEEE Trans. on Speech and Audio Proc.*, vol. 8, pp. 555–566, Sept. 2000.

[13] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. on Speech and Audio Proc.*, vol. 9, Mar. 2001.

[14] J.-L. Gauvain and C.-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. on Speech and Audio Proc.*, vol. 2, no. 2, April 1994.

[15] T. A. Myrvoll, K. K. Paliwal and T. Svendsen, "Fast Adaptation using Constrained Affine Transformations with Hierarchical Priors", Eurospeech, 2001.

[16] B. Launay, O. Siohan, A. C. Surendran, and C.-H. Lee, "Towards knowledge-based features for HMM based large vocabulary automatic speech recognition," in *Proc. ICASSP*, 2002.

[17] D. Ellis and M. J. R. Gomez, "Investigations into Tandem acoustic modeling for the Aurora task," in *Proc. of EuroSpeech*, 2001.

[18] M. Afify and O. Siohan, "Sequential noise estimation with optimal forgetting for robust speech recognition," in *Proc. ICASSP*, May 2001.