Telephone Band LVCSR for Hearing-Impaired Users

Ea-Ee Jan, Raimo Bakis, Fu-Hua Liu, Michael Picheny

IBM Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY
email: ejan,bakis,fhl,picheny@watson.ibm.com

ABSTRACT

Large vocabulary automatic speech recognition might assist hearing impaired telephone users by displaying a transcription of the incoming side of the conversation, but the system would have to achieve sufficient accuracy on conversational-style, telephone-bandwidth speech. We describe our development work toward such a system. This work comprised three phases: Experiments with clean data filtered to 200-3500Hz, experiments with real telephone data, and language model development. In the first phase, the speaker independent error rate was reduced from 25% to 12% by using MLLT, increasing the number of cepstral components from 9 to 13, and increasing the number of Gaussians from 30,000 to 120,000. The resulting system, however, performed less well on actual telephony, producing an error rate of 28.4%. By additional adaptation and the use of an LDA and CDCN combination, the error rate was reduced to 19.1%. Speaker adaptation reduces the error rate to 10.96%. These results were obtained with read speech. To explore the language-model requirements in a more realistic situation, we collected some conversational speech with an arrangement in which one participant could not hear the conversation but only saw recognizer output on a screen. We found that a mixture of language models, one derived from the Switchboard corpus and the other from prepared texts, resulted in approximately 10% fewer errors than either model alone.

I. Introduction

Advanced speech recognition technologies and affordable computation make it feasible to explore automatic speech recognition as an aid for hearing impaired people conversing over the telephone. Just imagine the following setup: the incoming telephony voice stream is fed into a PC-based speech recognition engine, and the transcription is displayed on a screen. The hearing impaired user looks at the displayed text and communicates via telephone with people without using today's cumbersome TDD or TTY system. Although current error rates on the ARPA Switchboard and Broadcast News tasks may seem too high for practical applications, we believe that a *speaker-dependent* telephony system can provide acceptable performance if users are cooperative and the task domain is suitably structured. In the future, additional technologies may help in this application, for example, speaker identification for automatic selection of speaker-dependent models, and speech synthesis for the impaired user who is not able to talk.

In our LVCSR system, words are represented as sequences of phones. Each phone, modeled by a three-state HMM, is further divided into 3 sub-phonetic units with context-dependent tying[2]. For each sub-phonetic unit, a decision tree is constructed from training data and the terminal nodes of the tree represent collections of instances of these classes grouped according to context. These context-dependent leaves are modeled by a mixture of Gaussian pdf's with diagonal covariance matrices. In our studies, the systems were built using approximately 2500 leaves with 30K and 120K Gaussians, approximately 12 and 60 Gaussians for each leaf, respectively. Different mel-cepstrum based feature spaces were used for the classifier, namely, 9 dimensional cepstra with normalized energy, 13 dimensional mel-cepstra (with C0) with their first and second order differences, and the same cepstra with several different feature space transformations. A weighted N-gram (bigram or trigram) is used to compute the language model probabilities. The signal processing of the feature space and language modeling will be discussed in detail in later sections.

The HMMs were trained by 40K in-house collected sentences, and the test set is an in-house office correspondence script, which includes 61 long sentences, with 1117 words. The wideband test set was collected through headset microphones, mainly ANC-500, and the telephony data was collected live through both local and long distance telephone networks, using several different telephone sets.

II. Acoustic Modeling on Desktop System

Prior to this study, a preliminary telephone band desktop system had been built using WSJ0 and WSJ1 training sets. In that system, the feature vectors were derived from 9 dimensional mel-cepstra with normalized energy, and their first and second order differences, using a Cepstrum Mean Normalization (CMN) scheme. The error rate on 10 speakers was approximately 25%.

A narrow band desktop system using clean, close-talking microphone data was then re-developed using in-house training data. This data was originally sampled at 16 KHz or higher, but was decimated to 8KHz and band-pass-filtered to 200-3500Hz.

Visual comparison of spectrograms reconstructed from 9 dimensional and 13 dimensional mel-cepstra suggests that there is useful information in the additional four cepstral components. When we tested both 9-dim CMN and 13-dim CMN in our recognizer, we found that the latter produced a relative

Sampling Rate	13 dim CMN			9 dim CMN
	16KHz	11KHz	8KHz	8KHz
Error Rate	12.7	14.6	16.8	18.0

Table 1: Error rates on different sampling rate.

error rate reduction of approximately 7%. (Table1) Although the computation of Gaussian densities in 13 dimensions is more expensive than in 9, the labeling accuracy with 13 dimensions is significantly better. This results in reduced searching time in the decoder, so that there is no significant increase in the overall computation time.

To study bandwidth effects, tests were conducted with sampling rates of 16kHz, 11kHz and 8 kHz, corresponding to Nyquist frequencies of 8 kHz, 5.5 kHz, and 4 kHz, respectively. Thirteen-dimensional cepstra (with C0) and their first and second order differences were used as feature vectors for these three systems. The same test set was downsmapled to each of these frequencies. The error rate increased from 12.7% in the 16KHz system to 16.8% in the 8KHz system. (Table 1) Approximately 2% absolute degradation was seen for each of these bandwidth reductions.[1]

Other signal processing schemes were evaluated in an effort to improve the performance. These are described below.

Linear Discriminant Analysis(LDA)[7] with nine concatenated frames of 13-dim cepstral vectors serving as input, rotated and reduced to a 39-dimensional output vector, resulted in a decrease of the error rate from 16.8% to 15.2%, which is a 9% relative improvement.

The intent of LDA is to transform the feature space to a coordinate system in which useful information is concentrated in a smaller number of coordinates and where the coordinate values are uncorrelated. The latter condition is helpful if the pdf's are to be modeled by Gaussians with diagonal covariance matrices. LDA analysis, however, looks only at the global average of the within-class covariance matrices, and ignores differences between them.

A more rigorous technique has been recently described for constructing a linear transformation to minimize the loss that results from constraining the covariance matrices to be diagonal [4]. We tested this Maximum Likelihood Linear Transformation (MLLT) using the same 117 dimensions (9 frames, 13 dim each) as input, and again transformed them to a 39-dimensional output vector. With this technique, the error rate dropped to 12.0%.

To explore performance as a function of the number of Gaussians, we built two systems, one with 120K Gaussians, (60-Gaussian mixture per leaf on average), and the other with 30K Gaussians (12 per leaf), using each of the above described signal precessing schemes. The results are summarized in Table2. As a consequence of all of these techniques, the error rate dropped from 25% to 12%.

	WSJ	*IBM Training data*			
Number of Gaussians	9dim CMN	9 dim CMN	13 dim CMN	13dim LDA	13dim MLLT
30K, SI	25	18	16.8	15.2	14.5
120K, SI		16.8	15.3	13	12
30K, SD		10.49	9.78	9.4	9.0

Table 2: Error rates on desktop, clean 8KHz Speaker Independent (SI) and Speaker Dependent (SD) systems.

Number of Gaussians	CMN	LDA	LDA+CDCN	telephony adapted
30K, SI	28.4	26.2	21.2	22.5
120K, SI	25.6		19.1	20.6
30K, SD	13.14	11.68	10.96	11.04

Table 3: Error rates on telephony Speaker Independent (SI) and Speaker Dependent (SD) systems.

The speaker dependent systems were then built and the results are listed at Table2.

III. Acoustic Modeling on Telephony Systems

Speech signals in telephone applications are inevitably vulnerable to the channel distortion and additive noise during transmission. Unlike the regular narrow-band applications, the distortion and noise in the telephone data can change from one recording to another significantly.

For baseline evaluation, the desktop system was tested with real telephony. (The test data was re-collected over telephone with the same number of speakers). As expected, considerable degradation occurred: the error rate jumped to 28.4%. The LDA system was not significantly better because the rotation matrix had been calculated from training data (clean speech) which is very different from telephony speech. To improve the performance, two approaches were tried. First, the system was adapted using a limited amount (1/7 of training data) of telephony data. The system was first adapted by MLLR[5] and then Gaussian smoothing, a scheme similar to MAP adaptation. The error rate came down to 20.6% (Table3). Feature-based signal processing techniques were then further explored.

A. The CDCN Algorithm

Since signal distortion and noise in telephone data can change from one recording to another, it can be very beneficial to enhance signals in feature space so that they appear to come from a more uniform acoustic condition. In this case, algorithms that can facilitate the simultaneous joint compensation for the effects of channel distortion and additive noise are highly desirable.

[1] **Special thanks to colleague Dr. Y. Gao who provided the wide band systems.**

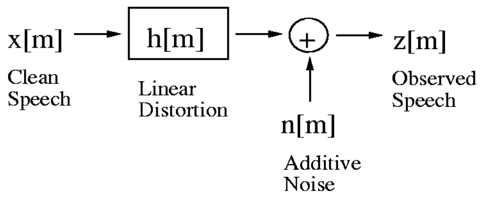

Figure 1: A model of environmental distortion used in CDCN

To this end, the Codeword-Dependent Cepstral Normalization (CDCN) algorithm developed at CMU [1] is explored in this paper. The CDCN algorithm assumes the model of environmental degradation shown in Fig. 1. The power spectrum can be written as in Eq. 1

$$P_z(\omega) = P_x(\omega)|H(\omega)|^2 + P_n(\omega) \tag{1}$$

and the corresponding cepstrum is written as

$$\mathbf{z} = \mathbf{x} + \mathbf{q} + \mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{q}) \tag{2}$$

where the perturbation vector $\mathbf{q} = IDFT[ln(|H(\omega)|^2)]$ represents the effect of linear filtering and the correction vector

$$\mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{q}) = IDFT[ln(1 + e^{\boldsymbol{DFT}[\mathbf{n}-\mathbf{q}-\mathbf{x}]})] \tag{3}$$

represents the joint effect of linear filter and additive noise.

Based on the structural knowledge of degradation model, CDCN attempts to solve two independent problems. The first problem is that of estimating the environmental parameters, $\mathbf{q}$ and $\mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{q})$, characterizing the contributions of additive noise and linear distortion. This is accomplished by using EM techniques to compute the ML estimation. The second problem is estimation of the un-corrupted observation vector $\mathbf{x}$ given the observed vector $\mathbf{z}$ and the estimated environmental parameters $\mathbf{q}$ and $\mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{q})$. MMSE parameter estimation is used for this task. In effect, these two operations determine the values of environmental parameters. When applied in an reverse fashion, they produce an ensemble of compensated vectors that best match, in the ML sense, the observed vectors in the testing environment to the locations of VQ codewords in the training environment, as shown in Eq.4

$$\hat{\mathbf{x}}_{\mathbf{i}} = \sum_{\mathbf{k}} \mathbf{f}_{\mathbf{i}}[\mathbf{k}] \left(\mathbf{z}_{\mathbf{i}} - \hat{\mathbf{q}} - \mathbf{r}^{(\mathbf{j})}[\mathbf{k}]\right) \tag{4}$$

where $f_i[k]$ is the weighting constant for Gaussian mixture k at frame i.

By applying CDCN, the acoustic features from telephony utterances were mapped to a more uniform space, and LDA was then applied to further optimize the feature space. The LDA transformation matrix was calculated using the training cepstra after CDCN mapping. The input and output acoustic feature vectors were kept as 117 and 39 dim, respectively. The CDCN+LDA system can dramatically reduce error rate from 28.4% to 19.1% (Table 3) without using a single telephony training sentence. In addition, a telephony system using more than 100 hours of telephony training data from different tasks was built later, and the results were comparable to this telephony CDCN+LDA system.

The speaker dependent systems were built and the results are listed at Table 3. Consistently, CDNC+LDA system has the best results.

The CDCN algorithm has the advantage that it does not require a priori knowledge of the testing environment. Although it is typically implemented on a sentence-by-sentence basis, it can be accomplished in a modeless fashion for real-time applications[6]. Since it does not assume acoustic similarity among the test data, CDCN is ideal for applications in which acoustic condition changes from sentence to sentence, such as in telephony applications.

IV. Language Modeling

The language model for the proposed application would ideally be trained and tested on spontaneous conversations recorded during actual use of the system. In the absence of such a complete system, we investigated the use of surrogate text sources to obtain an approximation to the desired language model. The designer of any language model for conversational speech is faced with the difficulty of obtaining sufficient amounts of representative text. Hundreds of millions of words are typically necessary to compute adequate statistics. Although machine-readable text corpora of this order of magnitude exist, they consist largely of news reports, literary works, legal, technical, and business correspondence and similar "written" sources. Currently available spontaneous conversation corpora tend to be much smaller. If a relatively small amount of text is available from a source that closely approximates the target application, and larger amounts are available from a source that is less similar to the target, then it may be possible to construct language models from both sources and combine them to obtain a better approximation to the target than either model alone [3]. Let p_1 be the probability for some word predicted by the first model, and p_2 the probability from the second model. The first suffers from random error because it is estimated from a small sample, the second suffers from bias because its sample was taken from a different population. A linear combination of the two,

$$p = wp_1 + (1 - w)p_2 \tag{5}$$

represents a smoothing of the noisy first model toward the less noisy but biased second model. Such a combination may produce a lower error rate than either original model alone. To optimize the value of the weight w, however, it is necessary to use an additional corpus which is more representative of the target application than is either of the other corpora. We started with a 2-million word corpus taken from the Switchboard Corpus, consisting of spontaneous telephone conversations on somewhat restricted topics. We smoothed this with a much larger corpus of over a hundred million words of prepared text including news, office correspondence, and literary works. Finally, to test the performance of the resulting mixture models, we generated a small amount of text in an arrangement intended to simulate the target application. A one-way tele-

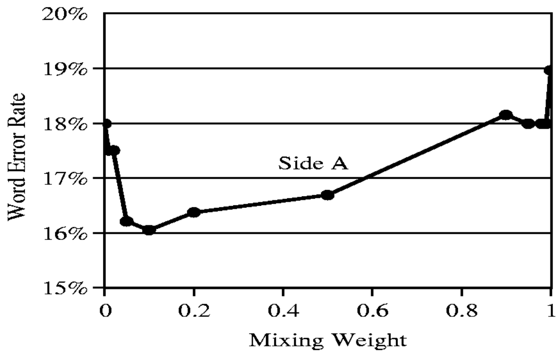

Figure 2: Error rate vs. language model mixing weight for spontaneous speech intended for a recognizer.

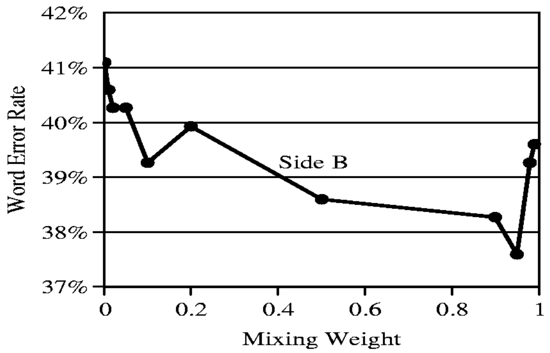

Figure 3: Error rate vs. language model mixing weight for spontaneous speech intended only for a human listener

phone connection enabled one participant, side A, to hear side B, but side B did not hear the speech of side A. That speech was, instead, fed into a speech recognition program and side B only saw the output of this program. Both sides of the conversation were recorded at a 4.5-KHz bandwidth for later off-line processing. The three corpora differed noticeably in style. For example, the probability of a sentence beginning with "I" was much higher in the Switchboard corpus than in the large prepared-text corpus, as was the frequency of conversational phrases such as "you know". The test corpus recorded in the simulated hearing-impaired situation contained requests for repetition when the recognizer made errors, which did not occur in either of the training corpora. The graphs below show the word error rates for the two sides of the conversation for various values of the mixing weight w. The value 1 means that only the Switchboard model was used, and 0 means that only the large, prepared-text model was used. Side A, the "hearing" participant, was aware that the other side, the "hearing-impaired" participant, was using a speech recognition system. Side A, therefore, tended to speak in a deliberate manner appropriate for automatic speech recognition. Side B, however, knew that the other participant listened directly, without the aid of a recognizer. Side B, therefore, spoke in a more casual style. Both sides were recorded, however, and then processed through a speech recognizer later off-line to get the results shown in the figures. The error rates clearly reflect the difference in speaking styles. In both cases, nevertheless, mixtures of the two language models produced lower error rates than either corpus alone.

It may be hoped that by collecting data more representative of the actual application, and possibly restricting the topics in some way, the error rates could be pushed down further.

V. Conclusion

Helping hearing-impaired telephone users by means of automatic speech recognition is a challenging opportunity. The conditions are adverse: limited bandwidth, variable channel characteristics, spontaneous conversational speech. Concerted application of state-of-the-art algorithms including adaptation to speaker and channel, together with a large set of acoustic prototypes, can bring acceptable accuracy within reach for read speech - we achieved a word error rate of 10.96%. It is also noteworthy that although other signal processing methods suffered significant loss when going from clean band-limited data to real telephone data, CDCN processing was able to compensate for telephone channel distortions, making it unnecessary to train on actual telephone data.

Although spontaneous conversation poses additional difficulties, these can be at least partly solved by improved language modeling. Design of the dialog to minimize LM perplexity, and cooperative users aware of the system's requirements should further improve accuracy.

References

[1] A. Acero, Acoustical and Environmental Robustness in Automatic Speech Recognition, Kluwer Academic Publishers, Boston, MA, 1993

[2] L.R. Bahl, et. al. "Performance of the IBM Large Vocabulary Continuous Speech Recognition System on the ARPA Wall Street Journal Task", 1995 ICASSP, pp 41-44

[3] F. Jelinek, B. Merialdo, S. Roukos and M Strauss, "A Dynamic Language Model for Speech Recognition", 1991 Proc. Speech and Natural Language DARPA Workshop, pp 293-295

[4] R. Gopinath, "Maximum Likelihood Modeling with Gaussian Distributions fo r Classification", 1998 ICASSP, pp 661-664

[5] C.J. Legetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMM's", Computer Speech and Language, vol 9, (2), pp 171-186

[6] F.H. Liu, A. Acero and R.M Stern, "Efficient Joint Compensation Of Speech For The Effects Of Additive Noise And Linear Filtering", Proc. of International Conference On Audio, Speech, And Signal Processing, 1992.

[7] H. Ney, R. Haeb-Umbach, B-H Tran and M. Oerder, "Improvement in Beam Search fro 10,000 Word Continuous Speech Recognition", Proc. ICASSP 92, pp I-9-12, 1992