# CONSTRUCTING TELEPHONE ACOUSTIC MODELS FROM A HIGH-QUALITY SPEECH CORPUS

*Mitchel Weintraub and Leonardo Neumeyer*

SRI International
Speech Research and Technology Program
Menlo Park, CA, 94025
USA

## ABSTRACT

In this paper we explore the effectiveness of constructing telephone acoustic models using a high-quality speech corpus. Results are presented for several front-end signal processing and feature mapping techniques. The algorithms were tested using SRI's DECIPHER™ speech recognition system [1-5] on several telephone corpora. The results show that (a) most of the performance loss when testing on telephone speech is due to the loss of information associated with the telephone bandwidth; (b) telephone-bandwidth systems trained with high-quality speech can outperform systems that are trained on telephone speech when tested on telephone speech; and (c) robust signal processing can allow speech recognizers to maintain performance when wide-bandwidth acoustic models are tested with telephone speech

## 1. INTRODUCTION

In many practical situations an automatic speech recognizer has to operate in several different but well-defined acoustic environments. For example, the same recognition task may be implemented using different microphones or transmission channels. In such situations it may not be practical to collect speech corpora to train the acoustic models of the recognizer for each acoustic environment. The research discussed in this paper focuses on how to modify acoustic models (such as those trained with high-quality Sennheiser recordings) for use over the telephone.

There are a number of possible approaches that we could take to modifying our high-quality acoustic models, including:

- Modify the front-end signal processing
- Use feature mapping
- Adapt the parameters of the acoustic models
- Play the high-quality data over the telephone
- Collect a new telephone corpus

To study the differences in speech recognition performance when training with high-quality speech compared to training with telephone-quality speech, we collected a pilot speech corpus focused on the Airline Travel Information Task (ATIS). This pilot corpus allowed us to compare different training and testing paradigms and evaluate recognition performance differences on a simultaneous test set consisting of both high-quality and telephone speech recordings. In addition, we report the results of experiments on the *Wall Street Journal* (WSJ) dictation corpus for simultaneous Sennheiser/telephone recordings when the acoustic models are trained with high-quality input. Finally, we compare speech recognition performance on the Switchboard Credit Card Corpus when training with either high-quality or telephone-quality acoustics.

Previous research by Chigier [6] has dealt with phoneme classification rates on the TIMIT and N-TIMIT corpora. Although he showed that the phone-classification error rates are lower for TIMIT than for N-TIMIT, the reason for this difference in performance was not determined. In this paper we show that the difference in performance between high-quality and telephone-quality speech is primarily due to the loss of information associated with the telephone bandwidth.

## 2. ATIS SIMULTANEOUS CORPUS

A corpus of both training and testing speech was collected using simultaneous recordings made from subjects wearing a Sennheiser HMD 414 microphone and holding a telephone handset. The speech from the telephone handset was transmitted over local telephone lines during data collection. The telephone speech was routed using a phone call that was made from an SRI internal line (interfaced to the SRI PBX) to an external Pacific-Bell line. The Pacific-Bell telephone line was interfaced to a Computerfone III which automatically answered the call and routed the speech audio to an Ariel analog-to-digital converter card.

Ten different telephone handsets were used. The telephones selected consisted of three carbon button telephones, two inexpensive Radio Shack telephones, and a variety of other common telephone types used in our lab. The same ten telephones were used for collecting both the training and testing sets. Only a single phone line was used for this corpus collection since the acoustic variation due to the telephone handset is one of the most important variables for telephone speech [7].

Thirteen male talkers who are familiar with speech recognition and the ATIS system were selected for this pilot data collection effort. Ten talkers were designated as training talkers, and three talkers were designated as the test set. Each of the training talkers read approximately 30 sentences for each of the ten different telephone handsets. The sentences that were used

for both training and testing prompts were selected at random from the ATIS MADCOW corpus [8]. The training set consisted of 3,000 simultaneous recordings of Sennheiser microphone and telephone speech. The test set consisted of 400 simultaneous recordings of Sennheiser and telephone speech.

## 3. ATIS EXPERIMENTAL RESULTS

The results obtained with this pilot corpus are shown in Table 1. The front-end signal processing consisted of 6 cepstral features in a tied-mixture acoustic model. The mean of each cepstral coefficient was removed on a per-sentence basis. The wide-bandwidth front-end signal processing consisted of a FFT-based filterbank using frequencies from 100-6400 Hz and 12 cepstral coefficients for the cepstral vector (C1-C12). The telephone-bandwidth front-end consisted of downsampling the speech data (16KHz samples/second to 8 KHz samples/second) followed by a FFT-based filterbank using frequencies 300-3300 Hz and 8 cepstral coefficients for the cepstral vectors (C1-C8). In both cases we remove the mean of each cepstral coefficient over the sentence. The number of cepstral coefficients for the delta and delta-delta cepstral vectors are the same as for the cepstral vector. The number of cepstral vectors used for each front-end had been optimized using other experimental corpora.

| Acoustic Model Training | | Test Set Word Error (%) | |
|---|---|---|---|
| Training Data | Front End Bandwidth | Sennheiser | Telephone |
| Sennheiser | Wide | 7.8 | 19.4 |
| Sennheiser | Telephone | 9.0 | 9.7 |
| Telephone | Telephone | 10.0 | 10.3 |

Table 1: Effect of Different Training and Front-End Bandwidth on Test Set Performance. Results are Word Error Rate on the 400 Sentence Simultaneous Test Set

We can see from Table 1 that there is a 15.4% decrease in performance when using a telephone front end (7.8% increases to 9.0% word error) and testing on Sennheiser data. This is due to the loss of information in reducing the bandwidth from 100-6400 Hz to 300-3300 Hz. However, when we are using a telephone front end, there is only a 7.8% increase in word error when testing on telephone speech compared to testing on Sennheiser speech (9.7% versus 9.0%). This is very surprising result, and we had expected a much bigger performance difference when Sennheiser models are tested on telephone speech acoustics.

Another surprising result shown in Table 1 is for acoustic models trained with telephone speech. For this experiment, performance is worse than when the acoustic models are trained with high-quality Sennheiser speech. (10.3% matched train and test on telephone compared to 9.7% train Sennheiser and test telephone).

In addition to these experiments, we performed a number of experiments to determine the robustness of a speech recognition system that uses wide-bandwidth acoustic models when tested with telephone speech. A MMSE spectral estimation algorithm was used to improve the robustness of these models. This algorithm is described in a companion paper for this conference [1] which we call Probabilistic Optimum Filtering (POF). This model is a piecewise linear transformation applied to the noisy speech observations; it constructs a minimum-mean square estimate of the clean speech features recorded using the Sennheiser microphone.

All POF mapping experiments use the simultaneous recordings from the training portion of the ATIS corpus described in Section 2 with the except of the “WSJ Robust POF15 mapping”, which used simultaneous recordings from the WSJ0 + WSJ1 corpus.

| Experiment | Telephone Word Error (%) |
|---|---|
| Baseline Zero-Mean Cepstrum | 19.4 |
| POF Mapping with Cepstrum | 9.4 |
| POF Mapping with Spectral SNR | 8.9 |
| POF Mapping with Cepstral SNR | 8.7 |
| WSJ Robust POF15 Mapping | 9.6 |

Table 2: Performance on ATIS Telephone Test Data using Wide-Bandwidth HMM Acoustic Models and Different Signal Processing Estimators

The results described in Table 2 show that the probabilistic mapping algorithm can be effectively used to reduce the error rate over the baseline signal processing algorithm. The error rate of the best-performing system on the telephone test set (8.7%) is only 11% higher than the error rate on the Sennheiser test (7.8%).

The error rate of the “POF mapping with Cepstral SNR” system (8.7%) is better than any telephone-bandwidth system listed in Table 1 (e.g. 9.7% trained with Sennheiser data or 10.3% trained with telephone data). There are several possible explanations for this. The telephone-bandwidth spectral analysis does not look at information outside of the frequencies from 300-3300 Hz. The information that is contained outside of the standard telephone bandwidth may be unreliable, thus causing performance to degrade when using the baseline processing algorithm which uses that information. The POF mapping algorithm is able to extract the useful information from 100-300 and 3300-6400 Hz.

A second explanation for why the “POF mapping with Cepstral SNR” outperforms the telephone-bandwidth spectral analysis is that it is exploiting correlations across both time and frequency to predict the missing information that is lost when

the speech is transmitted through the telephone network. In addition, the difference between the "POF Mapping with Cepstrum" and the "POF Mapping with Cepstral SNR" algorithm is that the latter mapping algorithm uses global information about the waveform. This global information is the estimate of the noise level.

## 4. WSJ EXPERIMENTAL RESULTS

A second set of experiments was performed on the *Wall Street Journal* (WSJ) Speech Corpus [9]. We evaluated our system on the 5000-word-recognition closed-vocabulary speaker-independent speech-recognition task Spoke S6: known microphone (telephone). This is a simultaneously recorded test set using both Sennheiser and an AT&T 712 telephone over local telephone lines.

The version of the DECIPHER speaker-independent continuous speech recognition system used for these experiments is based on a progressive-search strategy [3] and continuous-density, GENONIC hidden Markov models (HMMs) [2]. Gender-dependent models are used in all passes. Gender selection is accomplished by selecting the gender with the higher recognition likelihood.

The acoustic models used by the HMM system were trained with 37,000 sentences of Sennheiser data from 280 speakers, a set officially designated as the WSJ0+WSJ1 many-speaker baseline training. A 5K closed-vocabulary back-off trigram language model provided by M.I.T. Lincoln Laboratory for the WSJ task was used. Two front-end analyses are compared in the experiments below: a wide-bandwidth front-end analysis and a telephone-bandwidth front-end analysis. Gender-dependent HMM acoustic models were constructed for each of the two front-ends used.

The front-end processing extracts one long spectral vector which consists of the following six feature components: cepstrum, energy and their first and second order derivatives. The dimensionality of this feature is 39 (13 * 3) for the wide-bandwidth spectral analysis and 27 (9 * 3) for the telephone-bandwidth spectral analysis. The cepstral features are computed from an FFT filterbank, and subsequent cepstral-mean normalization on a sentence by sentence basis is performed.

Before using wide-bandwidth context-dependent genonic HMMs, a robust estimate of the Sennheiser cepstral parameters is computed using Probabilistic Optimum Filtering [1]. The robust front-end analysis is designed for an unknown microphone condition. The POF mapping algorithm estimates are conditioned on the noisy cepstral observations. Separate mappings are trained for each of the 14 microphones in the baseline WSJ0+WSJ1 si_tr_s stereo training. When the default no-transformation zero-mean cepstra are included, this makes a total of 15 estimated feature streams. These feature streams are computed on each test waveform, and the two feature streams with the highest likelihoods (using a simplified HMM for scoring the features) are averaged together. In all cases the first and second delta parameters are computed on these estimated cepstral values.

| Front-End Bandwidth | Signal Processing | Test Set | Word Error (%) |
|---|---|---|---|
| Wide | Standard | Sennheiser | 5.8 |
| Telephone | Standard | Sennheiser | 9.6 |
| Telephone | Standard | Telephone | 10.9 |
| Wide | Robust POF15 Cepstral Mapping | Telephone | 11.9 |

Table 3: Performance on the Aug 1993 WSJ Spoke S6 Development Test Set for Simultaneous Sennheiser/Telephone Recordings

The results in Table 3 show that most of the loss in performance between recognizing on high-quality Sennheiser recordings and on local telephone speech is due to the loss of information outside the telephone bandwidth. There is an increase in the word-error rate of 66% when testing on Sennheiser recordings with a wide-bandwidth analysis (5.8%) compared to testing with a telephone-bandwidth analysis (9.6%).

The loss in performance when switching from Sennheiser recordings to telephone recordings is small in comparison to the loss of information due to bandwidth restrictions. There is a 14% increase in the word-error rate when testing on the Sennheiser recordings (9.6%) compared to testing on the AT&T telephone recordings (10.9%).

Note that the performance using the "Robust POF15 Cepstral Mapping" with wide-bandwidth HMM acoustic models performs at almost the same level as a telephone-bandwidth HMM analysis (11.9 versus 10.9). This robust signal processing is able to maintain a high level of performance even when faced with dramatically different acoustic input.

In the ATIS experimental results described in Section 3, there was only an increase in word-error rate of 15% when switching from a wide-bandwidth analysis to a telephone-bandwidth analysis. However, in this experiment, we observed a 66% increase in the word-error rate when switching from the wide-bandwidth to the telephone-bandwidth conditions. There are several reasons for this difference. The first reason is due to the difference in tasks: the WSJ task has a larger vocabulary, contains more acoustically confusing words, and has a higher perplexity language model than the ATIS task. The second reason is that we used much better HMM acoustic models for the WSJ task (Genonic models) than for the ATIS task (tied-mixture models) and these differences between conditions are more noticeable when the word-error rates are lower.

The word-error rate for the WSJ Nov. 1993 evaluation test set was 8.8% on the Sennheiser microphone and 13.1% on the telephone handset using telephone-bandwidth acoustic models.

## 5. SWITCHBOARD CREDIT-CARD EXPERIMENTAL RESULTS

In many cases it is not possible to collect a simultaneous database consisting of Sennheiser recordings along with the desired type of recording. This is the case for a realistic telephone speech database.

A third set of experiments was performed on the Switchboard Credit Card task [10]. These experiments were performed at a workshop for digital analysis techniques of speech signals hosted by the CAIP center at Rutgers. The test-set used at the CAIP workshop consists of sentences extracted from continuous conversations between two talkers. These conversations were recorded digitally over long distance telephone lines, and is a 167 sentence subset of the Switchboard Corpus. The baseline language model for this workshop was provided by BBN.

Telephone-bandwidth phonetically tied-mixture HMM acoustic models were trained using the WSJ0 speaker-independent (84 talker) *Wall Street Journal* (WSJ) database [9] which was recorded using a high-quality Sennheiser microphone.

| Training Data | Word Error (%) |
|---|---|
| Credit Card | 68.1 |
| WSJ0 SI-Many Corpus | 71.5 |
| Credit Card Models Booted from WSJ0 Models | 67.1 |

Table 4: Word Error for Spontaneous Conversational Speech over Long Distance Telephone Lines

The word-error rates in this test set are very high and other speech recognition error rates at the conference were similar. We hypothesize that this is due to the conversational nature of the speech used for testing and not due to the acoustics of the test set. Note that the error rates when trained using 1100 credit-card telephone waveforms (68.1%) are very similar to those obtained when trained with 7000 WSJ0 high-quality waveforms (71.5%). In addition, the WSJ0 acoustic models can be used to improve the quality of the telephone acoustic models (67.1%) when they are used as the initial seed models.

## 6. CONCLUSIONS

We have presented speech-recognition results on three separate corpora: an ATIS test set consisting of simultaneous high-quality and telephone-quality recordings, a WSJ test set consisting of simultaneous high-quality and telephone-quality recordings, and a Switchboard test set consisting of sentences extracted from conversations over long-distance telephone lines.

We have shown that:

- Most of the performance loss in converting wide-bandwidth models to telephone speech models is due to the loss of information associated with the telephone bandwidth.
- It is possible to construct acoustic models for telephone speech using a high-quality speech corpus with only a minor increase in recognition word-error rate.
- A telephone-bandwidth system trained with high-quality speech can outperform a system that is trained on telephone speech even when tested on telephone speech.
- The variability introduced by the telephone handset does not degrade speech recognition performance.
- Robust signal processing can be designed to maintain speech recognition performance using wide-bandwidth HMM models with a telephone-bandwidth test set.

## ACKNOWLEDGEMENTS

The authors thank John Butzberger for helping to set up the ATIS experiments, and Vassilios Digalakis for providing Genonic HMM models and helping with the telephone-bandwidth experiments.

This research was supported by a grant, NSF IRI-9014829, from the National Science Foundation (NSF). It was also supported by the Advanced Research Projects Agency (ARPA) under Contracts ONR N00014-93-C-0142 and ONR N00014-92-C-0154.

## REFERENCES

1. L. Neumeyer, and M. Weintraub, “Probabilistic Optimum Filtering for Robust Speech Recognition,” 1994 IEEE ICASSP.
2. V. Digalakis and H. Murveit, “GENONES: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer,” 1994 IEEE ICASSP.
3. H. Murveit, J. Butzberger, V. Digalakis, and M. Weintraub, “Large-Vocabulary Dictation Using SRI’s DECIPHER™ Speech Recognition System: Progressive Search Techniques,” 1993 IEEE ICASSP, pp. II319-II322.
4. H. Murveit, J. Butzberger, and M. Weintraub, “Performance of SRI’s DECIPHER™ Speech Recognition System on DARPA’s CSR Task,” 1992 DARPA Speech and Natural Language Workshop Proceedings, pp. 410-414.
5. Murveit, H., J. Butzberger, and M. Weintraub, “Reduced Channel Dependence for Speech Recognition,” 1992 DARPA Speech and Natural Language Workshop Proceedings, pp. 280-284.
6. B. Chigier, “Phonetic Classification on Wide-Band and Telephone Quality Speech,” 1992 DARPA Speech and Natural Language Workshop Proceedings, pp. 291-295.
7. D. Kahn and A. Gnanadesikan, “Experiments in Speech Recognition over the Telephone Network,” 1986 IEEE ICASSP, pp. 729-732.
8. MADCOW, “Multi-Site Data Collection for a Spoken Language Corpus,” 1992 DARPA Speech and Natural Language Workshop Proceedings, pp. 7-14.
9. G. Doddington, “CSR Corpus Development,” 1992 DARPA SLS Workshop, pp. 363-366.
10. J.J. Godfrey, E.C. Holliman, and J.McDaniel, “SWITCHBOARD: Telephone Speech Corpus for Research and Development,” 1992 IEEE ICASSP, pp. I-517-I-520.

# CONSTRUCTING TELEPHONE ACOUSTIC MODELS FROM A HIGH-QUALITY SPEECH CORPUS

*Mitchel Weintraub and Leonardo Neumeyer*

SRI International
Speech Research and Technology Program
Menlo Park, CA, 94025
USA

## ABSTRACT

In this paper we explore the effectiveness of constructing telephone acoustic models using a high-quality speech corpus. Results are presented for several front-end signal processing and feature mapping techniques. The algorithms were tested using SRI's DECIPHER™ speech recognition system [1-5] on several telephone corpora. The results show that (a) most of the performance loss when testing on telephone speech is due to the loss of information associated with the telephone bandwidth; (b) telephone-bandwidth systems trained with high-quality speech can outperform systems that are trained on telephone speech when tested on telephone speech; and (c) robust signal processing can allow speech recognizers to maintain performance when wide-bandwidth acoustic models are tested with telephone speech

## 1. INTRODUCTION

In many practical situations an automatic speech recognizer has to operate in several different but well-defined acoustic environments. For example, the same recognition task may be implemented using different microphones or transmission channels. In such situations it may not be practical to collect speech corpora to train the acoustic models of the recognizer for each acoustic environment. The research discussed in this paper focuses on how to modify acoustic models (such as those trained with high-quality Sennheiser recordings) for use over the telephone.

There are a number of possible approaches that we could take to modifying our high-quality acoustic models, including:

- Modify the front-end signal processing
- Use feature mapping
- Adapt the parameters of the acoustic models
- Play the high-quality data over the telephone
- Collect a new telephone corpus

To study the differences in speech recognition performance when training with high-quality speech compared to training with telephone-quality speech, we collected a pilot speech corpus focused on the Airline Travel Information Task (ATIS). This pilot corpus allowed us to compare different training and testing paradigms and evaluate recognition performance differences on a simultaneous test set consisting of both high-quality and telephone speech recordings. In addition, we report the results of experiments on the *Wall Street Journal* (WSJ) dictation corpus for simultaneous Sennheiser/telephone recordings when the acoustic models are trained with high-quality input. Finally, we compare speech recognition performance on the Switchboard Credit Card Corpus when training with either high-quality or telephone-quality acoustics.

Previous research by Chigier [6] has dealt with phoneme classification rates on the TIMIT and N-TIMIT corpora. Although he showed that the phone-classification error rates are lower for TIMIT than for N-TIMIT, the reason for this difference in performance was not determined. In this paper we show that the difference in performance between high-quality and telephone-quality speech is primarily due to the loss of information associated with the telephone bandwidth.

## 2. ATIS SIMULTANEOUS CORPUS

A corpus of both training and testing speech was collected using simultaneous recordings made from subjects wearing a Sennheiser HMD 414 microphone and holding a telephone handset. The speech from the telephone handset was transmitted over local telephone lines during data collection. The telephone speech was routed using a phone call that was made from an SRI internal line (interfaced to the SRI PBX) to an external Pacific-Bell line. The Pacific-Bell telephone line was interfaced to a Computerfone III which automatically answered the call and routed the speech audio to an Ariel analog-to-digital converter card.

Ten different telephone handsets were used. The telephones selected consisted of three carbon button telephones, two inexpensive Radio Shack telephones, and a variety of other common telephone types used in our lab. The same ten telephones were used for collecting both the training and testing sets. Only a single phone line was used for this corpus collection since the acoustic variation due to the telephone handset is one of the most important variables for telephone speech [7].

Thirteen male talkers who are familiar with speech recognition and the ATIS system were selected for this pilot data collection effort. Ten talkers were designated as training talkers, and three talkers were designated as the test set. Each of the training talkers read approximately 30 sentences for each of the ten different telephone handsets. The sentences that were used

for both training and testing prompts were selected at random from the ATIS MADCOW corpus [8]. The training set consisted of 3,000 simultaneous recordings of Sennheiser microphone and telephone speech. The test set consisted of 400 simultaneous recordings of Sennheiser and telephone speech.

## 3. ATIS EXPERIMENTAL RESULTS

The results obtained with this pilot corpus are shown in Table 1. The front-end signal processing consisted of 6 cepstral features in a tied-mixture acoustic model. The mean of each cepstral coefficient was removed on a per-sentence basis. The wide-bandwidth front-end signal processing consisted of a FFT-based filterbank using frequencies from 100-6400 Hz and 12 cepstral coefficients for the cepstral vector (C1-C12). The telephone-bandwidth front-end consisted of downsampling the speech data (16KHz samples/second to 8 KHz samples/second) followed by a FFT-based filterbank using frequencies 300-3300 Hz and 8 cepstral coefficients for the cepstral vectors (C1-C8). In both cases we remove the mean of each cepstral coefficient over the sentence. The number of cepstral coefficients for the delta and delta-delta cepstral vectors are the same as for the cepstral vector. The number of cepstral vectors used for each front-end had been optimized using other experimental corpora.

| Acoustic Model Training | | Test Set Word Error (%) | |
|---|---|---|---|
| Training Data | Front End Bandwidth | Sennheiser | Telephone |
| Sennheiser | Wide | 7.8 | 19.4 |
| Sennheiser | Telephone | 9.0 | 9.7 |
| Telephone | Telephone | 10.0 | 10.3 |

Table 1: Effect of Different Training and Front-End Bandwidth on Test Set Performance. Results are Word Error Rate on the 400 Sentence Simultaneous Test Set

We can see from Table 1 that there is a 15.4% decrease in performance when using a telephone front end (7.8% increases to 9.0% word error) and testing on Sennheiser data. This is due to the loss of information in reducing the bandwidth from 100-6400 Hz to 300-3300 Hz. However, when we are using a telephone front end, there is only a 7.8% increase in word error when testing on telephone speech compared to testing on Sennheiser speech (9.7% versus 9.0%). This is very surprising result, and we had expected a much bigger performance difference when Sennheiser models are tested on telephone speech acoustics.

Another surprising result shown in Table 1 is for acoustic models trained with telephone speech. For this experiment, performance is worse than when the acoustic models are trained with high-quality Sennheiser speech. (10.3% matched train and test on telephone compared to 9.7% train Sennheiser and test telephone).

In addition to these experiments, we performed a number of experiments to determine the robustness of a speech recognition system that uses wide-bandwidth acoustic models when tested with telephone speech. A MMSE spectral estimation algorithm was used to improve the robustness of these models. This algorithm is described in a companion paper for this conference [1] which we call Probabilistic Optimum Filtering (POF). This model is a piecewise linear transformation applied to the noisy speech observations; it constructs a minimum-mean square estimate of the clean speech features recorded using the Sennheiser microphone.

All POF mapping experiments use the simultaneous recordings from the training portion of the ATIS corpus described in Section 2 with the except of the “WSJ Robust POF15 mapping”, which used simultaneous recordings from the WSJ0 + WSJ1 corpus.

| Experiment | Telephone Word Error (%) |
|---|---|
| Baseline Zero-Mean Cepstrum | 19.4 |
| POF Mapping with Cepstrum | 9.4 |
| POF Mapping with Spectral SNR | 8.9 |
| POF Mapping with Cepstral SNR | 8.7 |
| WSJ Robust POF15 Mapping | 9.6 |

Table 2: Performance on ATIS Telephone Test Data using Wide-Bandwidth HMM Acoustic Models and Different Signal Processing Estimators

The results described in Table 2 show that the probabilistic mapping algorithm can be effectively used to reduce the error rate over the baseline signal processing algorithm. The error rate of the best-performing system on the telephone test set (8.7%) is only 11% higher than the error rate on the Sennheiser test (7.8%).

The error rate of the “POF mapping with Cepstral SNR” system (8.7%) is better than any telephone-bandwidth system listed in Table 1 (e.g. 9.7% trained with Sennheiser data or 10.3% trained with telephone data). There are several possible explanations for this. The telephone-bandwidth spectral analysis does not look at information outside of the frequencies from 300-3300 Hz. The information that is contained outside of the standard telephone bandwidth may be unreliable, thus causing performance to degrade when using the baseline processing algorithm which uses that information. The POF mapping algorithm is able to extract the useful information from 100-300 and 3300-6400 Hz.

A second explanation for why the “POF mapping with Cepstral SNR” outperforms the telephone-bandwidth spectral analysis is that it is exploiting correlations across both time and frequency to predict the missing information that is lost when

the speech is transmitted through the telephone network. In addition, the difference between the "POF Mapping with Cepstrum" and the "POF Mapping with Cepstral SNR" algorithm is that the latter mapping algorithm uses global information about the waveform. This global information is the estimate of the noise level.

## 4. WSJ EXPERIMENTAL RESULTS

A second set of experiments was performed on the *Wall Street Journal* (WSJ) Speech Corpus [9]. We evaluated our system on the 5000-word-recognition closed-vocabulary speaker-independent speech-recognition task Spoke S6: known microphone (telephone). This is a simultaneously recorded test set using both Sennheiser and an AT&T 712 telephone over local telephone lines.

The version of the DECIPHER speaker-independent continuous speech recognition system used for these experiments is based on a progressive-search strategy [3] and continuous-density, GENONIC hidden Markov models (HMMs) [2]. Gender-dependent models are used in all passes. Gender selection is accomplished by selecting the gender with the higher recognition likelihood.

The acoustic models used by the HMM system were trained with 37,000 sentences of Sennheiser data from 280 speakers, a set officially designated as the WSJ0+WSJ1 many-speaker baseline training. A 5K closed-vocabulary back-off trigram language model provided by M.I.T. Lincoln Laboratory for the WSJ task was used. Two front-end analyses are compared in the experiments below: a wide-bandwidth front-end analysis and a telephone-bandwidth front-end analysis. Gender-dependent HMM acoustic models were constructed for each of the two front-ends used.

The front-end processing extracts one long spectral vector which consists of the following six feature components: cepstrum, energy and their first and second order derivatives. The dimensionality of this feature is 39 (13 * 3) for the wide-bandwidth spectral analysis and 27 (9 * 3) for the telephone-bandwidth spectral analysis. The cepstral features are computed from an FFT filterbank, and subsequent cepstral-mean normalization on a sentence by sentence basis is performed.

Before using wide-bandwidth context-dependent genonic HMMs, a robust estimate of the Sennheiser cepstral parameters is computed using Probabilistic Optimum Filtering [1]. The robust front-end analysis is designed for an unknown microphone condition. The POF mapping algorithm estimates are conditioned on the noisy cepstral observations. Separate mappings are trained for each of the 14 microphones in the baseline WSJ0+WSJ1 si_tr_s stereo training. When the default no-transformation zero-mean cepstra are included, this makes a total of 15 estimated feature streams. These feature streams are computed on each test waveform, and the two feature streams with the highest likelihoods (using a simplified HMM for scoring the features) are averaged together. In all cases the first and second delta parameters are computed on these estimated cepstral values.

| Front-End Bandwidth | Signal Processing | Test Set | Word Error (%) |
|---|---|---|---|
| Wide | Standard | Sennheiser | 5.8 |
| Telephone | Standard | Sennheiser | 9.6 |
| Telephone | Standard | Telephone | 10.9 |
| Wide | Robust POF15 Cepstral Mapping | Telephone | 11.9 |

Table 3: Performance on the Aug 1993 WSJ Spoke S6 Development Test Set for Simultaneous Sennheiser/Telephone Recordings

The results in Table 3 show that most of the loss in performance between recognizing on high-quality Sennheiser recordings and on local telephone speech is due to the loss of information outside the telephone bandwidth. There is an increase in the word-error rate of 66% when testing on Sennheiser recordings with a wide-bandwidth analysis (5.8%) compared to testing with a telephone-bandwidth analysis (9.6%).

The loss in performance when switching from Sennheiser recordings to telephone recordings is small in comparison to the loss of information due to bandwidth restrictions. There is a 14% increase in the word-error rate when testing on the Sennheiser recordings (9.6%) compared to testing on the AT&T telephone recordings (10.9%).

Note that the performance using the "Robust POF15 Cepstral Mapping" with wide-bandwidth HMM acoustic models performs at almost the same level as a telephone-bandwidth HMM analysis (11.9 versus 10.9). This robust signal processing is able to maintain a high level of performance even when faced with dramatically different acoustic input.

In the ATIS experimental results described in Section 3, there was only an increase in word-error rate of 15% when switching from a wide-bandwidth analysis to a telephone-bandwidth analysis. However, in this experiment, we observed a 66% increase in the word-error rate when switching from the wide-bandwidth to the telephone-bandwidth conditions. There are several reasons for this difference. The first reason is due to the difference in tasks: the WSJ task has a larger vocabulary, contains more acoustically confusing words, and has a higher perplexity language model than the ATIS task. The second reason is that we used much better HMM acoustic models for the WSJ task (Genonic models) than for the ATIS task (tied-mixture models) and these differences between conditions are more noticeable when the word-error rates are lower.

The word-error rate for the WSJ Nov. 1993 evaluation test set was 8.8% on the Sennheiser microphone and 13.1% on the telephone handset using telephone-bandwidth acoustic models.

## 5. SWITCHBOARD CREDIT-CARD EXPERIMENTAL RESULTS

In many cases it is not possible to collect a simultaneous database consisting of Sennheiser recordings along with the desired type of recording. This is the case for a realistic telephone speech database.

A third set of experiments was performed on the Switchboard Credit Card task [10]. These experiments were performed at a workshop for digital analysis techniques of speech signals hosted by the CAIP center at Rutgers. The test-set used at the CAIP workshop consists of sentences extracted from continuous conversations between two talkers. These conversations were recorded digitally over long distance telephone lines, and is a 167 sentence subset of the Switchboard Corpus. The baseline language model for this workshop was provided by BBN.

Telephone-bandwidth phonetically tied-mixture HMM acoustic models were trained using the WSJ0 speaker-independent (84 talker) *Wall Street Journal* (WSJ) database [9] which was recorded using a high-quality Sennheiser microphone.

| Training Data | Word Error (%) |
|---|---|
| Credit Card | 68.1 |
| WSJ0 SI-Many Corpus | 71.5 |
| Credit Card Models Booted from WSJ0 Models | 67.1 |

Table 4: Word Error for Spontaneous Conversational Speech over Long Distance Telephone Lines

The word-error rates in this test set are very high and other speech recognition error rates at the conference were similar. We hypothesize that this is due to the conversational nature of the speech used for testing and not due to the acoustics of the test set. Note that the error rates when trained using 1100 credit-card telephone waveforms (68.1%) are very similar to those obtained when trained with 7000 WSJ0 high-quality waveforms (71.5%). In addition, the WSJ0 acoustic models can be used to improve the quality of the telephone acoustic models (67.1%) when they are used as the initial seed models.

## 6. CONCLUSIONS

We have presented speech-recognition results on three separate corpora: an ATIS test set consisting of simultaneous high-quality and telephone-quality recordings, a WSJ test set consisting of simultaneous high-quality and telephone-quality recordings, and a Switchboard test set consisting of sentences extracted from conversations over long-distance telephone lines.

We have shown that:

- Most of the performance loss in converting wide-bandwidth models to telephone speech models is due to the loss of information associated with the telephone bandwidth.
- It is possible to construct acoustic models for telephone speech using a high-quality speech corpus with only a minor increase in recognition word-error rate.
- A telephone-bandwidth system trained with high-quality speech can outperform a system that is trained on telephone speech even when tested on telephone speech.
- The variability introduced by the telephone handset does not degrade speech recognition performance.
- Robust signal processing can be designed to maintain speech recognition performance using wide-bandwidth HMM models with a telephone-bandwidth test set.

## ACKNOWLEDGEMENTS

The authors thank John Butzberger for helping to set up the ATIS experiments, and Vassilios Digalakis for providing Genonic HMM models and helping with the telephone-bandwidth experiments.

This research was supported by a grant, NSF IRI-9014829, from the National Science Foundation (NSF). It was also supported by the Advanced Research Projects Agency (ARPA) under Contracts ONR N00014-93-C-0142 and ONR N00014-92-C-0154.

## REFERENCES

1. L. Neumeyer, and M. Weintraub, “Probabilistic Optimum Filtering for Robust Speech Recognition,” 1994 IEEE ICASSP.
2. V. Digalakis and H. Murveit, “GENONES: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer,” 1994 IEEE ICASSP.
3. H. Murveit, J. Butzberger, V. Digalakis, and M. Weintraub, “Large-Vocabulary Dictation Using SRI’s DECIPHER™ Speech Recognition System: Progressive Search Techniques,” 1993 IEEE ICASSP, pp. II319-II322.
4. H. Murveit, J. Butzberger, and M. Weintraub, “Performance of SRI’s DECIPHER™ Speech Recognition System on DARPA’s CSR Task,” 1992 DARPA Speech and Natural Language Workshop Proceedings, pp. 410-414.
5. Murveit, H., J. Butzberger, and M. Weintraub, “Reduced Channel Dependence for Speech Recognition,” 1992 DARPA Speech and Natural Language Workshop Proceedings, pp. 280-284.
6. B. Chigier, “Phonetic Classification on Wide-Band and Telephone Quality Speech,” 1992 DARPA Speech and Natural Language Workshop Proceedings, pp. 291-295.
7. D. Kahn and A. Gnanadesikan, “Experiments in Speech Recognition over the Telephone Network,” 1986 IEEE ICASSP, pp. 729-732.
8. MADCOW, “Multi-Site Data Collection for a Spoken Language Corpus,” 1992 DARPA Speech and Natural Language Workshop Proceedings, pp. 7-14.
9. G. Doddington, “CSR Corpus Development,” 1992 DARPA SLS Workshop, pp. 363-366.
10. J.J. Godfrey, E.C. Holliman, and J.McDaniel, “SWITCHBOARD: Telephone Speech Corpus for Research and Development,” 1992 IEEE ICASSP, pp. I-517-I-520.