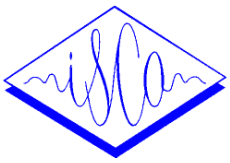

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION OF READ SPEECH OVER CELLULAR AND LANDLINE NETWORKS

Ashwin Rao, Bob Roth, Venkatesh Nagesha, Don McAllaster, Natalie Liberman, Larry Gillick

Dragon Systems, a Lernout & Hauspie company
320 Nevada Street, Newton, MA 02460, USA

ABSTRACT

We report results of large vocabulary continuous speech recognition (LVCSR) experiments, conducted using speech data read over cellular and landline phones. Specifically, we compare (using stereo recordings) the speaker-independent and speaker-adapted recognition word error rates (WERs) measured over cellular and landline networks, with those measured using a close-talking noise-canceling headset microphone, which serves as a baseline. A test set consisting of speech data recorded by 25 speakers is used; each speaker providing test and adaptation data. We use acoustic models trained from relatively high-quality training data and an interpolated trigram language model. Some insights into the relative degradation in WERs over telephone networks are also provided by examining the recognition error rates for bandlimited and coded microphone speech.

1. INTRODUCTION

Several results have been reported on the performance of speech recognition systems operating in telephony environments. Many early works have addressed speech recognition over landline networks [1]. More recently, there has been growing interest in extending speech recognition capabilities to handle cellular networks as well [6]. Researchers have focused on speech recognition tasks ranging from conversational telephone speech recognition [2] (e.g., the Switchboard Corpus) to highly constrained domain, large vocabulary, voice-enabled applications like stock quotes and call centers [3, 4]. Unfortunately, there seems to be limited knowledge of the difficulty involved in a large vocabulary task of transcribing speech over a (cellular or landline) telephone as opposed to a traditional close-talking microphone.

Our goal in this paper is to investigate the primary differences in recognition performance, when a recognizer is presented with speech data recorded over a cellular/landline telephone network compared to its close-talking microphone counterpart. To do so, we conduct experiments on in-house data, using Dragon's LVCSR system. In contrast to earlier published works, our experiments employ a test set that consists of speech data recorded by 25 speakers, each reading different scripts that were sampled from a large variety of topics of general interest with varying degrees of difficulties. The paper is organized as follows. The database description is provided in Section 2. In Section 3, we provide a brief overview of the particular LVCSR system that is used for experiments described in this paper. Finally, results are provided in Section 4, followed by Discussions and Conclusion in Section 5 and Section 6 respectively.

2. DESCRIPTION OF DATA COLLECTED

Each speaker recorded stereo data at the offices of Dragon Systems. In a first session, recordings were made over a cellular phone, while in a second session (recorded by most of the speakers), a landline phone was used. The reference channel in both sessions was a close-talking, noise-canceling headset microphone. The reference channel's data stream was recorded using a PC-based sound-card, while a Dialogic card (D-21A) was used for recording the telephone data. The cellular phone we used was a Nokia-2180 digital cellular phone using Bell Atlantic Mobile's code-division-multiple-access (CDMA) service. The landline phone was a Meridian telephone operating in a fixed network. In order to record data in a controlled fashion, a single example of each phone was employed for all the recordings. As a further control, we used the same room for all our recordings. The room's choice was influenced by many considerations that included size, background noise, reverberation, and proximity to a window (to get better cellular phone reception). However, each speaker made separate phone calls for the test and adaptation data, assuring that they constituted an independent sampling of the network.

For the cellular phone session, the data collection resulted in about 300 test utterances and a similar number of adaptation utterances for each of the 25 speakers (12 Males + 13 Females). Only a subset of these 25 speakers (9 Males + 11 Females) recorded the second session using the landline phone. The microphone data was sampled at 11 kHz (using a 16-bit linear PCM format) and the telephone data was sampled at 8 kHz (using an 8-bit μ-law format). Finally, the scripts that we handed speakers, were sampled from a large variety of topics of general interest with varying degrees of difficulties.

3. DRAGON'S LVCSR SYSTEM

We use a front-end that produces 24 dimensional feature vectors, which in turn are derived from a set of 36 feature parameters (12 Mel-cepstral + 12 delta + 12 delta-delta coefficients) using a Linear Discriminant Analysis (LDA) technique [7]. A standard cepstral mean subtraction is employed for channel normalization, and speaker normalization is performed by frequency warping [9], both during training and testing. We use different speaker independent (SI) acoustic models (AMs) for recognizing microphone and (cellular or landline) telephone speech. Both AMs are trained from a single large corpus that consists of more than 100 hours of high-quality microphone data using the same processing, except that the microphone models are trained from 5.5 kHz bandwidth data, while the telephone models are trained from 4 kHz bandwidth data. Finally, the speaker-adapted (SA) models are constructed using a transformation-based adaptation

technique [8]. All experiments employ an interpolated trigram language model with an out-of-vocabulary (OOV) rate of approximately 1.6%.

4. EXPERIMENTAL RESULTS

AMs	MIC1-WER	CELL-WER
SI	19.0%	31.0%
SA	16.0%	23.4%
SA-UnSupervised	16.5%	24.9%
AMs	MIC2-WER	TEL-WER
SI	18.9%	23.9%
SA	16.0%	19.3%
SA-UnSupervised	16.4%	20.3%

Table 1: WERs for recognition of noise-canceling microphone, cellular phone (CELL) and landline telephone (TEL) speech. Compared to the microphone WERs, the SI and SA WERs for CELL degrade by 63% and 46% respectively. For TEL, we see that the SI and SA WERs degrade by 27% and 21% respectively.

In this section, we first provide results of speech recognition experiments conducted on the database described in Section 2. The average WERs are reported in Table 1. MIC1 refers to microphone speech from the first session where a cellular phone (CELL) was recorded simultaneously. Similarly, MIC2 refers to the microphone speech from the second session where a landline phone (TEL) was used. Notice that the SI and SA WERs for the cellular phone degrade by 63% and 46% respectively, compared to the WERs for the reference microphone's speech. On the other hand, we see that the SI and SA WERs for the landline telephone degrade by only 27% and 21% respectively. In Table 1, we also provide results of performing unsupervised speaker adaptation, meaning making use of the recognizer, in conjunction with the baseline SI AMs, to generate transcripts for the adaptation data and using the latter to transform the SI AMs. One might worry that such unsupervised adaptation might suffer from the fact that the transcriptions created by the recognizer and used for adaptation are highly errorful (transcription WERs were around 25%, 40%, and 30% for microphone, cellular phone, and landline phone data respectively). However, it can be seen that for both cellular phone and landline telephone speech, almost 80% of the total improvement due to supervised adaptation can be obtained by doing unsupervised adaptation. We also experimented with doing unsupervised adaptation on the test data followed by a re-recognition of the same data (transcription-mode adaptation), which produced WERs of 17.3%, 26.4%, 17%, and 21.3% for MIC1, CELL, MIC2, and TEL respectively. These are higher than the WERs for unsupervised adaptation on the adaptation data, partly because twice as much adaptation data, compared to test data, is used in the adaptation computation.

4.1. Jackknifing for producing Channel-Adapted AMs

Recall that the telephony AMs that we use are essentially built from bandlimited high-quality data. Training AMs with large amounts of real-world cellular phone and landline phone data may seem tempting. However, we now present some empirical evidence that seems to indicate that training AMs from telephone data may not necessarily yield desirable improvements.

AMs	CELL-WER	TEL-WER
SI-J	28.0%	22.6%
SA-J	23.5%	19.4%

Table 2: WERs for recognition of CELL and TEL speech, using AMs built by jackknifing. Using approximately 11 hours of real cellular phone data to channel-adapt the SI AMs, reduces the baseline SI WER by 3 points. Using roughly 9 hours of real landline phone data, reduces WER for landline telephone speech by 1.3 points. For both the cellular and landline conditions, the WERs after speaker adaptation are very similar to the baseline SA WERs. Further, the relative WERs between SI-J WERs and the corresponding SI WERs (from Table 1), are similar to the relative WERs between SA-J WERs and their corresponding SA WERs (from Table 1).

We designed some jackknifing experiments, using the database described in Section 2, to study this issue.

The jackknifing experiments for recognizing the k-th speaker's cellular phone speech proceed as follows. We first do transformation-based adaptation of the SI AMs using *all other speakers'* cellular phone adaptation data. We denote this AM as SI-J. Next, we further transform the SI-J AM by performing speaker adaptation using the k-th speaker's cellular phone adaptation data. Let us denote this as SA-J. We then test the k-th speaker's cellular phone speech using the SI-J and SA-J AMs. This procedure is repeated for all the speakers and for their landline data as well. The resulting average WERs are listed in Table 2.

By comparing the SI-J WERs from Table 2 with the SI WERs from Table 1, we see the following. By using approximately 11 hours of real cellular phone data to channel-adapt the SI AM, the baseline WER is reduced by 3 points. In addition, by using roughly 9 hours of real landline phone data, the WER for landline telephone speech is reduced by 1.3 points. Interestingly, for both types of phones, the WERs after speaker adaptation are very similar to the baseline SA WERs, as seen by comparing the SA-J WERs from Table 2 with the corresponding SA WERs from Table 1. Further on, notice that the relative WERs between SI-J WERs from Table 2 (for CELL and TEL) and the SI WERs from Table 1 (for MIC1 and MIC2 respectively) are similar to the relative WERs between SA-J WERs from Table 2 (for CELL and TEL) and the SA WERs from Table 1 (for MIC1 and MIC2 respectively). This seems to indicate that the absolute SI-J WERs (for the cellular and landline telephones) may be further reduced, only by marginal amounts, by training from additional within domain telephone data. However, based on the above experiments, one may argue that training AMs from real-world telephone data may improve the SI performance, but may not necessarily alter the WERs if, in any event, one were to adapt using the particular cellular/landline phone.

4.2. Varying Amount of Adaptation Data

We study here the effect of varying the amount of adaptation data. This may be of particular interest in telephone speech recognition applications, where the adaptation time may play an important role. In Figure 1, we have displayed the WERs as a function of the number of utterances used for SA. The solid lines

correspond to results of supervised adaptation experiments. The dashed lines refer to results of doing unsupervised adaptation. The results for MIC2 are similar to those of MIC1. Clearly, at least 50% of the improvement due to full adaptation (using 300 utterances) may be achieved by adapting with only 50 utterances; adapting with only 100 utterances seems to provide almost $3/4$ of the improvement. Also, for small number of adaptation utterances, the differences, in terms of WERs, between unsupervised and supervised adaptations are small.

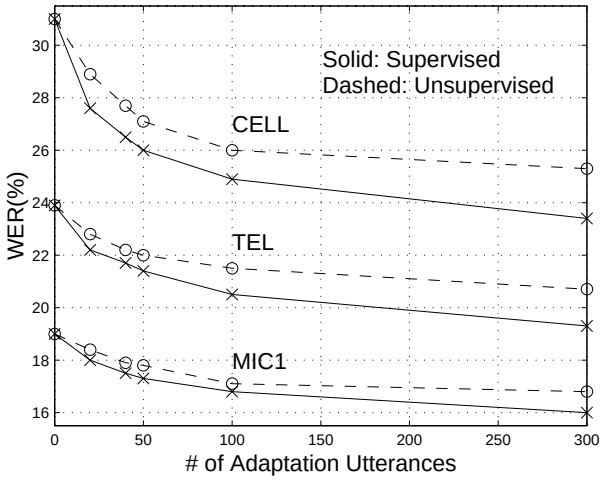

Figure 1: WERs measured while recognizing with models that were speaker-adapted using only a portion of the adaptation data.

4.3. Effects of Bandwidths on WER

Recall that the WERs in Table 1 compare the recognition performance on telephone data with that on 5.5 kHz bandwidth microphone data. We report results of experimenting with microphone data with varying bandwidths in Table 3. As before, the tests using microphone data with 0–5.5 kHz bandwidth use an AM trained from 0–5.5 kHz bandwidth. However, to test MIC1 and MIC2 data with bandwidths of 0–4 kHz and 260 Hz–3.4 kHz, we use the telephony AM that was trained from data that was bandlimited to 0–4 kHz. For the SA experiment, we adapted the SI AM using bandlimited adaptation data.

As can be seen, for a typical telephone bandwidth (with a typical range of 260 Hz – 3.4 kHz), the baseline SI and SA WERs increase by around 19% and 11% respectively. As a comment, the extra degradation in SI WER for the telephone bandwidth, compared to its SA WER, may be due to a mismatch between the training and test data, since we use an AM trained from data whose spectra are limited to 0–4 kHz, and we test on data with a 260 Hz–3.4 kHz bandwidth. In general though, compared to the WER for microphone data with a telephone bandwidth of 260 Hz – 3.4 kHz, the SI and SA WERs for the cellular phone are degraded by 37% and 32% respectively, and those for the landline telephone are degraded by 6% and 8% respectively.

4.4. Effects of CODECs on WER

Typically, voice transmission over a cellular network calls for coding of the message signal (speech in our case); coding is typically performed to compress the modulating signal so as to increase the overall throughput of the channel. Broadly speaking,

Bandwidth (Hz)	MIC1-WER (%)		MIC2-WER (%)	
	SI	SA	SI	SA
0–5500 (Baseline)	19.0	16.0	18.9	16.0
0–4000	20.5	17.2	20.5	17.6
260–3400	22.6	17.8	22.5	17.9

Table 3: WERs for recognizing microphone data, with varied bandwidths. Comparing the WERs for the microphone speech bandlimited to reflect a typical telephone bandwidth of 260 Hz – 3.4 kHz, the SI and SA relative-WERs for the cellular phone are 37% and 32% respectively, and those for the landline telephone are 6% and 8% respectively.

wireless telephony coding standards may be classified into two categories: the ones standardized for Global System for Mobile Communications (commonly referred to as GSM), along with variants like Time-Division-Multiple-Access (TDMA), which is being used widely in Europe and, to a lesser extent, in the United States, and the coders employed in networks based on the CDMA technology (prevalent mainly in the U.S.). Within these classes, several codecs exist that primarily differ in their transmission bit-rate capabilities (ranging from low bit-rate coders operating at 2.4–4.8 kilobits-per-second (kbps) to codecs operating at medium (8-16 kbps) and high bit rates (more than 30 kbps)). More recently, there have been systems that also employ what is called as variable bit-rate codecs (wherein the bit rate is varied based on whether the modulating speech is voiced, unvoiced, silence, etc.). In [5], some effects of a sample of codecs on small vocabulary speech recognition performance were reported. In this section, we consider two such typical hybrid codecs (operating at medium bit rates) and evaluate their performance, in the context of large vocabulary continuous speech recognition.

In Table 4 we have listed the results. The experiments were carried out on the microphone channel speech data. Specifically, the signal samples were bandlimited, encoded using the particular codec and then decoded back to yield the samples that were fed to the recognizer. For the adaptation experiments, the adaptation data was signal-processed based on the coding scheme and subsequently used for adaptation.

First, we notice that μ-law encoding (which is typical for landline transmission) does not worsen the WERs. On the other hand, the LPC based encoders like the GSM (at 13.2 kbps) and the CELP (at 9.6 kbps) seem to have a significant effect on the WER. For instance, at a telephone bandwidth of 260 Hz–3.4 kHz, the effect of bandwidth limitation coupled with the LPC codecs is observed to degrade the baseline SI WERs by almost 20%, and the SA WERs by more than 10%. We have also provided WERs for recognition of bandlimited microphone speech that has undergone "tandeming" (multiple coding schemes performed sequentially) in Table 4. The results are similar to those previously reported [5]. If tandeming of codecs occurs for the entire duration of cellular phone calls, then even as few as 2 codecs in tandem could significantly increase the baseline WER.

5. DISCUSSION

Our knowledge of the precise coding scheme employed in the Nokia cellular phone we used for our recordings is limited.

Bandwidth (Hz)	Coding	MIC2-WER (%)	
		SI	SA
0–4000	16-bit LIN-PCM	20.5	17.6
0–4000	8-bit μ-Law	20.3	17.5
Bandwidth (Hz)	Coding	MIC1-WER (%)	
		SI	SA
0–4000	16-bit LIN-PCM	20.5	17.2
0–4000	GSM	22.0	18.6
0-4000	CELP	22.5	18.8
260-3400	GSM	24.1	19.6
260-3400	CELP	24.7	19.1
0–4000	GSM+GSM	24.8	20.1
0–4000	CELP+CELP	27.3	21.5
0–4000	GSM+CELP	25.6	20.7
0–4000	CELP+GSM	25.2	20.4

Table 4: WERs for recognizing bandlimited and compressed microphone speech. For the telephone bandwidth of 260–3400 Hz, the GSM (at 13.2 kbps) and CELP (at 9.6 kbps) codecs seem to degrade the baseline SI WER by almost 20%, and the SA WER by slightly more than 10%.

However, comparing the 19.1% SA WER for telephonized and CELP-coded microphone data with the cellular phone's 23.4% SA WER, it appears that there is an additional degradation of 20–25% from unknown sources which remains to be explained. This additional degradation remains, even though our data has been collected in reasonably controlled environments, and does not reflect the effects of long-distance calls, multiple locations, different handsets, doppler spread, and other sources of variation [10] typically inherent in cellular networks.

Compared to microphone WERs, the relative degradation in WER for a landline phone seems significantly smaller than that for a cellular phone. Certain characteristics of our test data may be worth listing: (1) the average signal-to-noise ratios (SNRs) measured for the microphone, landline, and cellular channels were 30, 28, and 26 dBs respectively, and (2) signal dropouts and crosstalks (characteristic of cellular networks) were detected (by manually listening to the audio) in less than 3% of the total number of cellular phone utterances. We do not believe that additive noise or other obvious sources of distortion account for a major portion of the additional degradation in WER for cellular networks. It may very well be that the effect of codecs on WER, operating in real-world environments, is more severe than what we measured using clean microphone data. One possible explanation for some of the unexplained degradation may be the spectral/temporal distortions introduced in the signal by cellular networks [10], for example due to fading (envelope and multipath), which are not satisfactorily captured by our acoustic models, even after adaptation. We are currently investigating these issues.

The relative WERs that we reported, comparing the recognition performance of microphone speech with telephone speech, may seem to be a reasonable measure. However, we must caution the reader that even the relative WERs may vary, based on the absolute reference WER and also on the specific task being considered. For instance, for a large vocabulary task in a highly constrained domain, we have observed that the SI WER on a landline telephone, relative to the SI WER on a reference close-talking headset microphone, degrades by less that 10%, while for the task addressed in this paper, the corresponding number is 27%.

6. CONCLUSION

In this paper, we showed that the SI and SA WERs for the cellular phone degrade by 63% and 46% respectively, relative to the WERs for reference microphone speech. On the other hand, the SI and SA WERs for the landline telephone speech degrade by only 27% and 21% respectively. We demonstrated that unsupervised adaptation yields almost 80% of the total improvement due to supervised adaptation. Our jackknifing experiments further demonstrated that training acoustic models from telephone speech may improve the SI performance, but may not necessarily alter the SA performance; possibly due to the fact that speaker adaptation techniques also adapt the acoustics to the particular channel. Experimental results on reduced adaptation data were presented, showing that almost $3/4$ of the improvement due to adaptation on 300 utterances, may be achieved by using only a third of the data. Finally, it was demonstrated that the combined effect of bandwidth limitation and coding, explains almost half of the degradation in WER for the cellular phone.

References

1. L. Neumeyer, V. Digalakis, and M. Weintraub, "Training Issues and Channel Equalization Techniques for the construction of Telephone Acoustic Models Using a High-Quality Speech Corpus", *IEEE Trans. Speech and Audio Proc.*, vol. 2, No. 4, Oct'94.
2. B. Peskin *et al.*, "Improvements in recognition of conversational telephone speech", *Proc. ICASSP'99*, pp. 53-56.
3. J. Bernstein, K. Taussig, and J. Godfrey, "Macrophone: An American English Telephone Speech Corpus for the Polyphone Project", *Proc. ICASSP'94*, pp. 81–84.
4. S. Das *et al.*, "Towards Robust Speech Recognition in the Telephony Network Environment - Cellular and Landline Conditions", *Proc. Eurospeech'99*.
5. B. Lilly and K. Paliwal, "Effect of Speech Coders on Speech Recognition Performance", *Proc. ICSLP'96*, pp. 2344–2347.
6. R. Haeb-Umbach, "Robust Speech Recognition for Wireless Networks and Mobile Telephony", *Proc. Eurospeech'97*, pp. 2427–2430.
7. M. J. Hunt *et al.*, "An Investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination", *Proc. of ICASSP'91*, pp. 881–884.
8. V. Nagesha and L. Gillick, "Studies in Transformation-Based Adaptation", *Proc. ICASSP'97*, pp. 1031–1034.
9. S. Wegmann *et al.*, "Speaker Normalization on Conversational Speech", *Proc. ICASSP'96*, pp. 339–341.
10. K. Feher, "Wireless Digital Communications", *Prentice-Hall PTR*, New Jersey.