

Effect of Speaking Style on LVCSR Performance

Mitch Weintraub, Kelsey Taussig, Kate Hunicke-Smith, Amy Snodgrass

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

ABSTRACT

SRI collected a corpus to study how spontaneous speech differs from other types of speech. The corpus was collected in two parts: (1) a spontaneous Switchboard-style conversation on an assigned topic, and (2) a reading session in which participants read transcripts of their conversation from part 1. Experiments were conducted on sentences with identical transcripts that varied in speaking style. The word-error rates varied from 29% (careful dictation) to 53% (spontaneous conversation) depending on the speaking style. These experiments show that speaking style is a dominant factor in determining the performance of large-vocabulary conversational speech recognition (LVCSR) systems.

1. INTRODUCTION

It is well known that speaking style has an effect on the performance of speech recognition systems. From a research point of view, the important questions are:

1. How can we quantitatively measure the changes in speech for different speaking styles? How can we model these changes? Can we identify the speaking style from a given observation?
2. What are the effects of changes in speech style on the performance of human and computer speech recognition?
3. How can we make computer speech recognition algorithms more robust to changes in speech style?

Recent research to address the first question has primarily focused on the Lombard effect for isolated words [Junqua '93]. This research showed that the Lombard effect changes many of the different acoustic parameters that are used to characterize speech (formants, bandwidths, duration, pitch, cepstral coefficients, ...).

There have been a few studies that focused on the phonological differences between spontaneous and read speech (fast, normal, and slow reading) [Bernstein et al. '86, Cohen et al. '87, Cohen '88]. These studies showed that when instructed to read faster, subjects mostly increased the rate by speeding up each segment that is spoken. This was in contrast to spontaneous speech, where a fast rate is accomplished by deleting phonemes.

One of the first efforts to quantify the impact of speech styles on speech recognition performance was undertaken by the

DARPA Robust Speech Recognition Effort. These studies [Rajasekaran et al. '86, Paul et al '86] focused on isolated and connected words from a small vocabulary from a large range of speech styles (normal, lombard, deliberate, loud, soft, shout, fast, angry), while under motion (twisting, lying on back) as well as different vibration conditions. These studies showed that automatic speech recognition systems had the highest error rates for fast, loud, angry, and shouting conditions.

Another set of experiments focused on changes in speech style when interacting with a computer to make airline flight reservations [Wade et al. '92, Butzberger et al. '92]. They found that word-error rates decreased as users interacted with the same SLS system over time. This effect was attributed to changes in speech style by the user.

There has also been several algorithmic approaches to compensating for different speech styles [Lippmann et al. '87, B.A. Hansen & Applebaum '90, J. Hansen '95]. These studies showed that performance could be improved by (1) training with speech collected under different conditions, (2) using first and second derivatives of the cepstral parameters, and (3) smoothing the variance of the model parameters.

None of the earlier work focused on natural conversational speech. The goal of these experiments was to determine how speaking style affects the word-error rate for LVCSR systems. Specifically, we wanted to contrast speech produced in a spontaneous conversational style with read speech. The read speech was collected by having subjects read transcripts of their previous conversations. In this way, the text of the input speech remained fixed (thereby eliminating any differences in performance due to the language model). This allows us to compare acoustic differences that are a result of differences in speaking style.

The organization of the paper is: data collection methodology is described in section 2, the LVCSR system is described in section 3, experimental results are presented section 4, and section 5 summarizes the results.

2. DATA COLLECTION

SRI collected the corpus described in this paper to study how spontaneous speech differs from other types of speech. The corpus was collected in two parts:

1. A spontaneous Switchboard-style conversation on an assigned topic, and
2. A reading session in which participants read transcripts of their own conversation.

During the reading session, the participants were instructed to read transcripts from their side of the conversation in two styles:

1. As if they were dictating to a computer
2. As if they were having a conversation.

All conversations were recorded both over the telephone and also as high-bandwidth 16-bit waveforms using Sennheiser microphones. The high-bandwidth 16-bit waveforms were recorded as using an SGI Indy with a Sennheiser HMD 410 microphone. The telephone waveforms were recorded over digital telephone lines using a Dialogic DTI101 with an AT&T 712 handset.

This corpus was collected in two portions:

1. In spring of 1995, SRI collected the spontaneous portion of the corpus. Thirty subjects participated; fifteen male and fifteen female. This portion was modelled as closely as possible after Texas Instruments' original Switchboard data collection [J. Godfrey et al. '92, on the web: http://www ldc.upenn.edu/ldc/catalog/html/speech_html/scr.html].
2. In summer of 1995 twenty of SRI's original thirty subjects came back to record read versions of their original spontaneous conversations.

2.1. Spontaneous Corpus Partition

Topics were taken from TI's original set of Switchboard topics. The fifteen topics used were: air pollution, buying a car, crime, ethics in government, exercise and fitness, family life, gun control, latin america, music, painting, pets, public education, space flight and exploration, universal public service, and wood-working.

The thirty subjects, 15 male, 15 female, were recruited locally. Paired subjects did not know each other (excepting pair #13), and did not meet face to face until after their conversation had taken place. This was done to keep the SRI collection as similar to the TI collection as possible. The thirty subjects were paired to provide 5 male-male pairs, 5 female-female, and 5 male-female pairs.

Members of the pairs were in separate, carpeted offices during the recording. Speech was recorded over two channels; telephone and high quality. Each subject wore a head mounted microphone with one earpad flipped up to accommodate a telephone handset. The head mounted microphone was a Sennheiser HMD-410 high quality noise-cancelling model connected to an SGI INDY through a Rane MS-1 pre-amplifier. The subject

could hear their conversational partner through the one Sennheiser headphone. The telephone was an AT&T model 712. All telephone data was recorded directly from T1 digital lines in 8-bit mu-law format using a Dialogic DTI101 recording system.

2.2. Read Corpus Partition

Twenty of the thirty original subjects, (11 males and 9 females), came back to record the read portion of the corpus.

SUBJECT INSTRUCTION:

Subjects were instructed in the proper use of the data collection tool, especially the hold-to-talk button, to avoid truncation of the waveforms.

Subjects read four distinct sets of data in the following order:

1. A set of 40 common (same for all speakers), phonetically balanced sentences
2. Approximately thirty sentences from North American business news articles
3. The text of their side of their previously recorded conversation, read in a dictation style
4. The text of their side of their previously recorded conversation read in a conversational style

The subjects were instructed to read the first three sets as if they were dictating to a computer, and to read the last set (read set #4) as if they were having an actual conversation. These instructions were given at the beginning of the appointment, and repeated before each set.

3. LVCSR SYSTEM DESCRIPTION

The SRI DECIPHER(TM) speaker-independent continuous speech recognition system is based on continuous-density, genonic hidden Markov models (HMMs) [Digalakis '94]. The system used a multi-pass recognition strategy [Murveit '93] with a vocabulary of 20K words, non-cross word acoustic models and a bigram language model.

The front end signal processing is based on an FFT-based spectral analysis every 10msec, which are integrated into 18 spectral bands from 300 to 3300 Hz, and are used to compute 9 cepstral coefficients (C1-C8) plus C0. The mean of the C1-C8 cepstral coefficients are removed over the whole conversation using all the sentences (and using all frames) on this side. From these 9 cepstral features (C0-C8), first and second derivatives over time are computed. These features are concatenated together to form a 27 dimensional cepstral vector, which is used as the single input feature of the recognition system.

All the speech segments on this side of the conversation are used to make the gender determination. A two state HMM with 256 mixtures is used to perform gender selection (1 state for male speech, 1 state for female speech). The feature for this classification is the cepstral vector C1-C8. The probabilities are accumulated across all sentences for this conversation side and a decision is made at the end of the conversation side.

After performing gender selection, a gender-dependent genonic HMM is used in a 2-pass recognition strategy (forward and backward passes) to generate word-lattices for that sentence using a bigram grammar. The initial 2-pass system used a bigram language model with explicit bigram transitions between words.

The acoustic models used for this task are non-crossword genonic acoustic models. The CMU 100K WSJ lexicon was used as a pronunciation dictionary. This pronunciation dictionary was augmented by adding the 100 most common words in the Switchboard corpus not in the CMU lexicon, plus all the non-speech sounds (e.g. [noise]) with pronunciations of a signal “reject” phone. All nonspeech sounds and all words without pronunciations were trained to use this single reject phone.

The male non-cross word acoustic models consisted of 7083 triphones and 2011 biphone models. Each of these phone models used 3-state linear topologies. These acoustic models were clustered into 918 different sets of 32 Gaussian mixtures. Each of the triphone and biphone models had it’s own 32 mixture weights on the shared Gaussian mixtures. The female non-cross word acoustic models consisted of 6670 triphones and 1975 biphone models. These models were clustered into 730 clusters of 32 Gaussian mixtures for the female system. Since these were non-cross word acoustic models, they did not extend across word boundaries. No word-specific models were used.

ACOUSTIC TRAINING:

The training data consisted of both Macrophone (http://www ldc upenn edu ldc/catalog/html/speech_html/macrophone.html) and Switchboard training data. The Macrophone data consisted of 21K male sentences and 24K female sentences. Of the Macrophone corpus, we used the TIMIT, ATIS, and WSJ sentences and a small percentage of the other sentences (date, time, numbers, ...). The speech from the Switchboard corpus used was the full training set. About 10 percent of the Switchboard training data was eliminated due to the lack of ability to generate training alignments with the hypothesized words. There were 91K Switchboard male sentences for a total of 112K male training sentences. There were 97K Switchboard female sentences, for a total of 121K training sentences.

GRAMMAR TRAINING:

For training the language model and selecting the vocabulary, we used the BBN segmented development test text. This text corresponds to the allowable Switchboard training partition filtered through the BBN segmentation algorithm. This text was further processed to remove punctuation and case distinction. In addition, all noise words were collapsed into a single [noise] tag. The training data therefore consisted of 248,151 sentences, with a total of 1,916,827 words.

There were 22,365 unique words in this training data and this constituted the basis for the language model. A back-off trigram language model was constructed, which consisted of 22365 unigrams, 309,264 bigrams, and 891,531 trigram transitions.

The intersection of the 22,456 words and the CMU lexicon produced a recognition vocabulary of 20,653 words. All the non-

speech words (e.g. the [noise] word) were added to the CMU lexicon with a pronunciation of the “reject” phone before this step. All other words that were not in the dictionary were deleted from the language model. This resulted in a bigram language model containing 305,016 (of the total 309,264) bigram transitions.

4. EXPERIMENTAL RESULTS

Recognition experiments were performed on all recordings with identical transcriptions in a 3 styles (spontaneous, read dictation, and read conversational). If a speaker did not read his original conversational speech exactly (word-for-word) in any of the read versions, the sentence was put into a separate category. For example, there may have been disfluencies in the original spontaneous speech that the talker was not able to reproduce exactly. There were 660 sentences with identical transcripts and 523 sentences with some difference in the transcripts.

The recognition results for all data with identical transcriptions in shown in the table below:

Speaking Style	Word Error
Read Dictation	28.8%
Read Conversational	37.6%
Spontaneous Conversation	52.6%

The recognition results for spontaneous utterances with non-identical recordings was 55.1%. These results demonstrate that style of speech in the dominant factor in recognition error rate.

5. SUMMARY

Experiments were conducted on sentences with identical transcripts that varied in speaking style. By keeping the LVCSR system fixed, and by keeping the transcripts fixed, we were able to focus on the style of speech production.

The more casual the speaking style, the higher the LVCSR word-error rate. Note that the two styles of reading (as if dictating to a computer or a conversational style) had significantly different error rates (29% and 38% respectively). Also note that when people thought that they were reading in a conversational style, the error rates were still much lower than actual conversational style (38% and 53% respectively).

These experiments show that speaking style is a dominant factor in determining the performance of large-vocabulary conversational speech recognition (LVCSR) systems. We did not try to control for other factors (e.g. speech rate).

These experiments lead us to believe that there is significant fluctuation in the acoustics realization of a word sequence that is not currently modeled by triphone-based GMM systems. The focus of future research will be to determine what the factors are that affect performance and how to incorporate these variables into LVCSR systems.

REFERENCES

1. J.C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," 1993 JASA Vol. 93(1), pp. 510-524.
2. J. Bernstein, G. Baldwin, M. Cohen, H. Murveit, and M. Weintraub, "Phonological Studies for Speech Recognition," Proceedings of DARPA Speech Recognition Workshop, February 19-20, 1986, pp. 41-48.
3. M. Cohen, G. Baldwin, J. Bernstein, H. Murveit, and M. Weintraub, "Studies for an Adaptive Recognition Lexicon," Proceedings of DARPA Speech Recognition Workshop, March 24-26, 1987, pp. 49-55.
4. M. Cohen, *Phonological Structures for Speech Recognition*, U.C. Berkeley Ph.D. thesis, 1989
5. P.K. Rajasekaran, and G.R. Doddington, "Robust Speech Recognition: Initial Results and Progress," Proceedings of DARPA Speech Recognition Workshop, February 19-20, 1986, pp. 73-80
6. D.B. Paul, R.P. Lippmann, Y. Chen, and C.J. Weinstein, "Robust HMM-Based Techniques for Recognition of Speech Produced Under Stress and in Noise," Proceedings of DARPA Speech Recognition Workshop, February 19-20, 1986, pp. 81-92.
7. E. Wade, E. Shriberg, and P. Price, "User Behaviors Affecting Speech Recognition," 1992 ICSLP, pp. 995-998.
8. J. Butzberger, H. Murveit, E. Shriberg and P. Price, "Spontaneous Speech Effects in Large Vocabulary Speech Recognition Applications," 1992 DARPA Speech and Natural Language Workshop, pp. 339-343.
9. R.P. Lippmann, E.A. Martin, D.B. Paul, "Multi-Style Training for Robust Isolated-Word Speech Recognition," Proceedings of DARPA Speech Recognition Workshop, March 24-26, 1987, pp. 96-99.
10. B.A. Hansen and T.H. Applebaum, "Robust speaker-independent word recognition using static, dynamic, and acceleration features: Experiments with Lombard and noisy speech," 1990 ICASSP, pp. 857-860.
11. J.H.L. Hansen, and M.A. Clements, "Source Generator Equalization and Enhancement of Spectral Properties for Robust Speech Recognition in Noise and Stress," 1995 IEEE Trans. Speech and Audio Processing, Vol. 3., No. 5, pp. 407-415.
12. M. Weintraub, H. Murveit, M. Cohen, P. Price, J. Bernstein, G. Baldwin, D. Bell, "Linguistic Constraints in Hidden Markov Model Based Speech Recognition," IEEE ICASSP 1989 pp. 699-702.
13. 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, pp. I537-I540.
14. H. Murveit, J. Butzberger, V. Digalakis and M. Weintraub, "Large-Vocabulary Dictation Using SRI's DECI-PHER(TM) Speech Recognition System: Progressive-Search Techniques," 1993 IEEE ICASSP, pp. II-319 - II-322.
15. J. J. Godfrey, E.C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," 1992 IEEE ICASSP, pp. 517-520.