MULTI-STYLE TRAINING FOR ROBUST ISOLATED-WORD SPEECH RECOGNITION*

Richard P. Lippmann, Edward A. Martin, Douglas B. Paul

Lincoln Laboratory, Massachusetts Institute of Technology
Lexington, Massachusetts 02173-0073

ABSTRACT

A new training procedure called multi-style training has been developed to improve performance when a recognizer is used under stress or in high noise but cannot be trained in these conditions. Instead of speaking normally during training, talkers use different, easily produced, talking styles. This technique was tested using a speech data base that included stress speech produced during a workload task and when intense noise was presented through earphones. A continuous-distribution talker-dependent Hidden Markov Model (HMM) recognizer was trained both normally (5 normally spoken tokens) and with multi-style training (one token each from normal, fast, clear, loud, and question-pitch talking styles). The average error rate under stress and normal conditions fell by more than a factor of two with multi-style training and the average error rate under conditions sampled during training fell by a factor of four.

INTRODUCTION

The performance of current recognition systems often degrades dramatically as a talker's speech characteristics change with time, when a talker is under normal levels of workload or psychological stress, and when a talker is in a high noise environment. New techniques to prevent this degradation have been developed and tested with a number of data bases including a new Lincoln stress-speech data base. In this paper we first review results obtained with this speech data base and then provide detailed information on the effects of multi-style training. Other papers in this proceedings describe discriminant analysis [1] and cepstral stress compensation [2] and present results obtained with another speech data base [3].

Lincoln Stress-Speech Data Base

The Lincoln stress-speech data base includes words spoken with eight talking styles (normal, slow, fast, soft, loud, clear enunciation, angry, question pitch) and under three stress conditions. A difficult motor-workload task [4] was used to create easy (cond50) and more difficult (cond70) workload stress conditions that emulate the type of workload stress experienced when driving a car or flying an airplane. A third stress condition was created by presenting 85 dB SPL of speech-shaped noise through earphones. This produces the so-called Lombard effect [8] where a talker speaks louder and often more clearly when in noise. This is the main cause for recognizer degradation in noise in situations where an acoustically-shielded close-talking microphone minimizes the effect of additive noise. The data base vocabu-lary contained 35 difficult aircraft words with acoustically similar subsets such as go, hello, oh, no, and zero. A total of 11,340 tokens were obtained from 9 male talkers during three sessions per talker spanning a four week period.

HMM Recognizer

The baseline continuous-distribution HMM recognizer described in [4] was used for all experiments. It is a left-to-right isolated-word recognizer with multivariate Gaussian distribu-tions and diagonal covariance matrices where observations consist of centisecond mel-scale cepstral parameters. Unless otherwise stated, all results were obtained using 10-node word models created using five training tokens per word with the forward-backward algorithm [5] and using the Viterbi algorithm [5] during recognition.

RESULTS WITH LINCOLN STRESS-SPEECH DATA BASE

Figure 1 presents an overview of results in rough chronological order obtained using a number of different techniques with the Lincoln stress-speech data base. The initial error rate, averaged over all conditions excluding the most difficult angry condition, was 17.5%. A similar high error rate was obtained with a new, high performance, commercial recognizer. Poor perfor-mance for the initial Lincoln system and the commercial system was caused by the difficult vocabulary and stress conditions and by the fact that only normally-spoken speech was used in training. The initial Lincoln recognizer was the baseline system with variance limiting [4] which limits the variance estimates obtained during forward-backward training to be above a specified lower limit. The high initial error rate was more than halved to 6.9% using multi-style training. In this case, the five tokens used during training were taken from the normal, fast, clear, loud, and question-pitch talking styles instead of only from the normal style. Multi-style training halved the

CH2396-0/87/0000-0705 $1.00

error rate with no increase in computation requirements.

The next large reduction in error rate (from 6.9% to 3.2%) was obtained by doubling the number of parameters used in the observation vector. The original vector of 16 cepstral parameters was supplemented with 16 additional differential parameters which were the differences between the current 16 parameters and the parameters computed 20 ms earlier. This differential parameter technique was also recently used by [6]. It reduces the error rate, but also doubles the recognition computation requirements. The next large decrease in error rate (from 3.2% to 1.6%) was obtained by using grand-variance estimates. Instead of estimating the variance of each of the 32 observation parameters separately for each node of every word model, the grand variance of each observation parameter was estimated once across all word models and all nodes during training. Using grand variances reduces the degradation in performance caused by using a statistical model that is too complex for the amount of training data. This result reinforces past results that demonstrate the necessity of matching the complexity of a model to the amount of training data [7]. Using grand variances halved the error rate while simultaneously decreasing recognition computation requirements. The final large reduction in error rate (1.6% to 1.0%) was obtained using the two-stage discriminant analysis system described in [1]. This system focuses attention on those parts of often confused words that are most different and reduces the error rate with only a slight increase in recognition computation requirements. The final system with a 1% error rate across many stress/style conditions is a usable, practical, robust recognizer that could be used for a variety of speech-recognition tasks.

EFFECTS OF MULTI-STYLE TRAINING

More details on the effects of multi-style training from the experiments described above are presented in Figs. 2 to 4. Figure 2 compares results with normal and multi-style training for the six novel conditions not sampled during training as well as for normally-spoken speech. These are representative results for the situation where a recognizer cannot be trained under live stress conditions. The percentage error rate averaged over all nine talkers is presented for normal speech, for speech spoken slowly, for the easy (cond50) and the more difficult (cond70) workload task, for soft speech, for speech produced in noise (Lombard) and for angry speech. Multi-style training reduces the error rate substantially for all conditions. The average error rate over all conditions fell by more than a factor of two from 20.7% to 9.8%. The drop in error rate is large (6.2% to 2.9%) even for normally spoken words and greatest for the Lombard and angry conditions.

Figure 3 shows the results when the recognizer was tested under the same conditions sampled during training. Here, the average error rate over all conditions fell by a factor of four from 18.4% to 4.6%. It should be noted that in these and other experiments, training word tokens were never used during testing.

Further experiments were performed to determine whether more effective subsets of five styles could be found and whether fewer than five different styles could provide large improvements. These experiments suggest that the five styles selected are more effective than other subsets of the eight styles in the stress-speech data base and that all five different styles are required for best performance with multi-style training. Further experiments have also been performed to explore the effects of multi-style training with more advanced HMM isolated-word talker-dependent recognizers. We have found that multi-style training always improves overall performance. For example, the error rate for an advanced recognizer with differential parameters, grand-variance estimates, 14 nodes, and five training tokens, drops from 3.2% to 1.4% with multi-style training.

One surprising result evident in Figs. 2 and 3 is that the error rate drops for normal speech when the recognizer is trained on non-normal training tokens. This is caused by day-to-day variability in normal speech as demonstrated in Fig. 4. Figure 4 presents the error rate with normal and multi-style training for normal speech recorded in the first, second, and third recording sessions. As can be seen, multi-style training and normal training produce similar results in session one, but multi-style training is superior in sessions two and three. These results demonstrate that multi-style training can compensate for variability in normal speech over time, and that five normal training tokens recorded in one session are less representative of normal tokens recorded one to three weeks later than five multi-style tokens.

DISCUSSION

Multi-style training improves performance for the novel stress conditions because: (1) the forward-backward training algorithm and statistical decoding focuses attention on spectral/temporal regions that are consistent across styles and (2) speech samples are presented during training that are similar to those that occur during testing. For example, loud speech is similar in many ways to speech produced under the Lombard condition. The improvement in performance with conditions sampled during training was greater than the improvement with novel untrained condition for this second reason.

A careful analysis of differences between word models obtained using normal and multi-style training and of recognizer confusions indicated that improvements are caused by two main mechanisms. First, estimates of the mean and variance of the cepstral parameters used in HMM word models are more representative of those observed during testing with multi-style training. This is illustrated in Fig. 5. The left side of this figure presents the difference

between multi-style and normally trained cepstral mean estimates and the right side presents the ratio of multi-style over normally trained cepstral variance estimates. Data are averaged over all talkers and all word models. As can be seen, the lower-order cepstral mean estimates are reduced with multi-style training. This compensates for spectral tilt (presumably caused by narrower glottal pulses) which is characteristic of much of the stress speech in the Lincoln data base. Variance estimates for lower-order cepstral coefficients are also higher with multi-style training. This weights these cepstral coefficients less heavily during recognition because they are more variable across stress and style conditions.

A second mechanism that leads to better performance with multi-style training is that word models are richer and provide a better description of perceptually important acoustic events that are present across talking styles. This mechanism was discovered by examining spectrograms created from normal and multi-style HMM word models for those models that caused major confusions. Spectrograms were created by plotting the average spectrum at each node with duration equal to the average node residency time. For example, Fig. 6. contains spectrograms generated from HMM word models for the word "break". The left spectrogram was generated from a normally-trained word model and the right one was generated from a multi-style model. Numbers indicate the HMM node number used to generate each spectra. In these and all experiments, the end nodes (nodes numbered 0 and 9 in Fig. 6) are anchors that match background noise. The large ticks in Fig. 6 are at 100 ms intervals, the lower curve plots overall energy, and the frequency scale extends to roughly 6 kHz. As can be seen, the multi-style word model contains the optional release for the final /k/ and provides a clearer description of formant transitions. Examination of many other word-model spectrograms showed that multi-style word models generally contain more of the important acoustic-phonetic cues used in spectrogram reading than normally-trained models.

SUMMARY

A new training procedure called multi-style training was developed and tested with a stress-speech data base. It improves performance substantially under stress and with different talking styles, and can be used when a recognizer cannot be trained under live stress conditions. It also improves performance under normal conditions by compensating for normal day-to-day speech variability.

REFERENCES

[1] E. A. Martin, R. P. Lippmann, and D. B. Paul, "Two-Stage Discriminant Analysis for Improved Isolated-Word Recognition," ICASSP'87, Dallas, TX, April 1987.

[2] Y. Chen, "Cepstral Domain Stress Compensation for Robust Speech Recognition," ICASSP'87, Dallas, TX, April 1987.

[3] D. B. Paul, "A Speaker-Stress Resistant HMM Isolated-Word Recognizer," ICASSP'87, Dallas, TX, April 1987.

[4] 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," Proc. Speech Tech 86, pp. 241-249, New York, NY, April 1986.

[5] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," BSTJ, 62, pp. 1035-1074, April 1983.

[6] E. L. Bocchieri, and G. R. Doddington, "Frame-Specific Statistical Features for Speaker Independent Speech Recognition," IEEE Trans. Acoust. Speech, and Signal Processing, ASSP-34, pp. 755-764, August 1986.

[7] G. F. Hughes, "On the Mean Accuracy of Statistical Pattern Recognizers," IEEE Trans. Info. Theory, IT-14, pp. 53-63, January 1968.

[8] E. Lombard, "Le Signe de l'Elevation de la Boix," Ann. Maladiers Oreille, Larynx, Nez, Pharynx, 37, 1911.

*This work was sponsored by the Defense Advanced Research Projects Agency.
The views expressed are those of the authors and do not reflect the official policy or position of the U.S. Government.

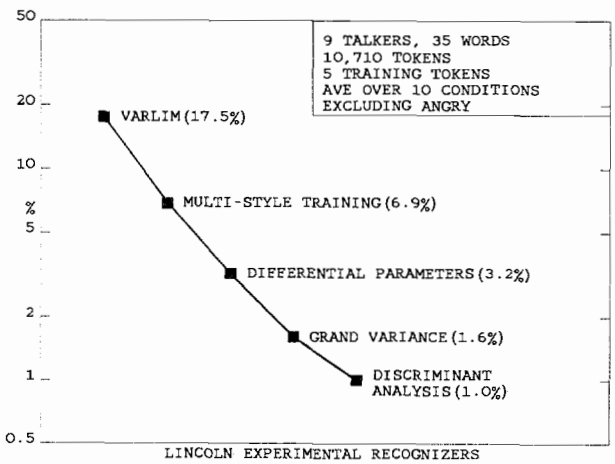

Fig. 1. Substitution errors with Lincoln stress-speech data base.

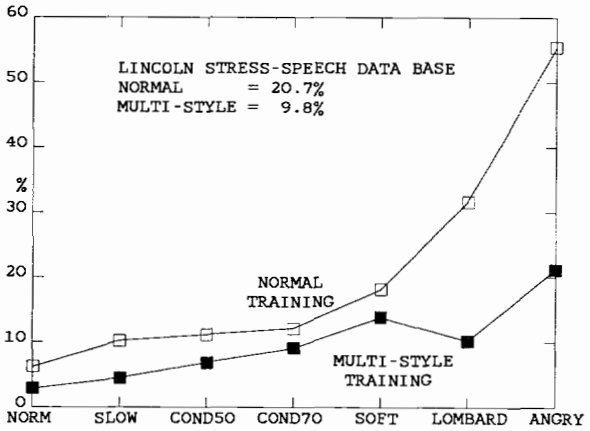

Fig. 2. Percent errors for novel untrained conditions.

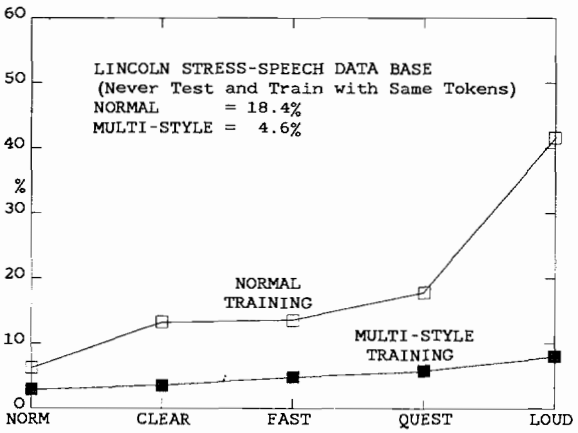

Fig. 3. Percent errors for conditions used during training.

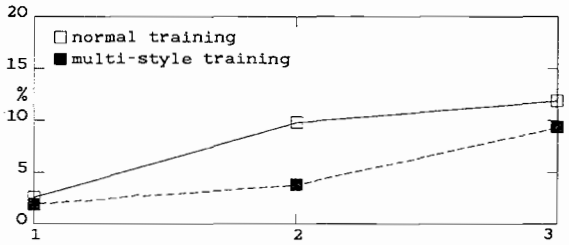

Fig. 4. Percent errors for normal speech across recording sessions.

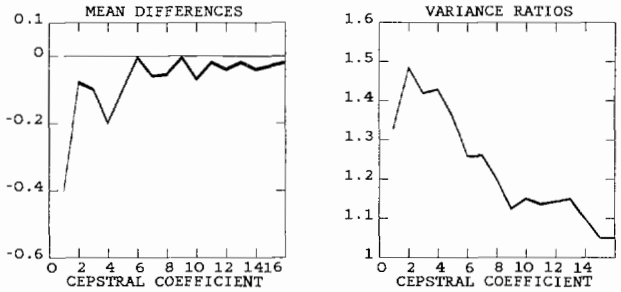

Fig. 5. Differences between means (A) and ratios of variances (B) from HMM models created using multi-style training and normal training.

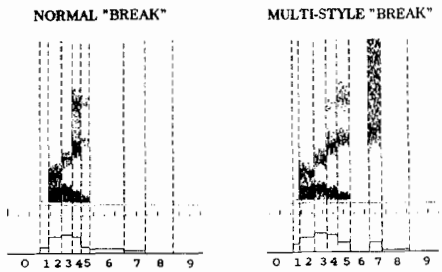

Fig. 6. Spectrograms of the word "break" created from normally trained HMM word model (A) and multi-style trained model (B).