MULTI-STREAM ASR TRAINED WITH HETEROGENEOUS REVERBERANT ENVIRONMENTS

Michael L. Shire

University of California at Berkeley
International Computer Science Institute
Berkeley, California USA
shire@icsi.berkeley.edu

ABSTRACT

A common problem with current automatic speech recognition (ASR) systems is that the performance degrades when it is presented with speech from a different acoustic environment than the one used during training. An important cause is that the feature distribution to which the ASR system is trained no longer matches that of a new environment. Reverberant environments can be especially harmful. In this work, we test a multi-stream system in which the constituent streams are each trained in separate acoustic environments. When training the acoustic modeling stages of the streams separately with clean data and heavily reverberated data, we find that that the combined system can improve the ASR performance with unseen reverberated test data.

1. INTRODUCTION

Despite many advances in ASR robustness, ASR systems can still perform abysmally when the test data and the data used to train the system are from different acoustic environments. The acoustic environment can significantly alter the distributions of the speech features. For example, room reverberation results in the time-translated signal energy to be added to itself, resulting in modulation energy that is "smeared" forward in time. When the effect is significant, the trained acoustic models are no longer accurate. Researchers have sometimes used model adaptation techniques or preprocessing techniques with robust properties to alleviate the problems due to a change in acoustic environment. As a complementary alternative, we choose instead to keep multiple system components that are independently trained to more than one acoustic environment. In our experiments, we employ multiple front-end acoustic modeling streams to estimate class posteriors that are then merged prior to decoding. One stream or set of streams are trained using the original clean data. The remaining stream or streams are trained in a heavily reverberated environment. In this manner, the ASR system has examples of the what the feature distribution looks like with data subjected to two extremes of reverberation.

2. EXPERIMENTAL SETUP

Experiments were performed using a hybrid artificial neural network - hidden Markov model system [2]. A simple three layer feed forward multi-layer perceptron (MLP) estimates the posterior probabilities of mono-phone class targets given the acoustic features. A subset of the OGI Numbers corpus [3] was used for recognition experiments. This corpus consisted of naturally spoken connected numbers recorded over the telephone and has a small vocabulary size of 32 words. The training set consisted of approximately 3 hours of speech while the development testing set contained about 1 hour of speech. A smaller third cross-validation (CV) set was used as hold-out data for early-stopping during the MLP training and was sometimes used for parameter tuning of the decoder. The MLPs used an input context of 9 frames of speech features. We used the *chronos* decoder [9] with a bigram grammar and phone models derived from phonetic transcriptions of the corpus.

The training and testing speech utterances were used in their original state (clean condition) and also modified with examples of reverberation. One set of examples was estimated from recordings in a variable echoic chamber composed of panels that could be placed in a highly reflective or absorbent state. Varying the percentage of panels that were in either state yielded room impulses with different reverberant characteristics. One example from this set consisted of light reverberation whose impulse response was estimated from recordings in a variable echoic chamber [10]. It had the quality of a small office with a reverberation time (T60) of 0.5 seconds and a direct-to-reverberant ratio (DTR) of -2 dB. It was used in preliminary tests on combinations while the remaining impulses were saved for final experiments. An additional reverberation example consisted of a heavy reverberation, whose impulse was constructed from recordings in a concrete basement hallway and has a T60 of 2.5 seconds and of DTR of -8 dB. Additional artificial examples of reverberation were constructed by simple modification of the heavy reverberation impulse to approximate reverberant environments with different characteristics, i.e DTR and T60. Reverberated data was applied artificially to the speech via convolution with the room impulse response.

3. SIMPLE COMBINATION RESULTS

There are a number of ways to combine probability estimates from several classifiers. These are usually based upon different independence assumptions [1]. A common method used by researchers is to average the logarithm of the probabilities,

$$P(q|x_1, x_2, \ldots, x_N) = \frac{\exp(\sum_{i=1}^{N} w_i \log P(q|x_i))}{\sum_{q'} \exp(\sum_{i=1}^{N} w_i \log P(q'|x_i))}$$

where $P(q|x_i)$ is the posterior probability of class q given features x of stream i, and the weights w_i sum to one. When the weights w_i are equal, this equation is equivalent to the normalized geometric mean of the posteriors. This method has similar properties to the product rule combination method [8] and has often produced

Submitted to ICASSP 2001 Salt Lake City, 2000-NOV-10

Do not distribute

MLP Combination	WER (%) clean test	light test	heavy test
clean MLP alone	6.4	27.0	68.5
heavy MLP alone	49.0	37.1	38.0
log-average	10.0	21.3 +	49.3
clean big MLP	6.0	26.2	68.9
heavy big MLP	49.2	37.2	36.3
clean+heavy big MLP	10.9	21.0	38.1 +

Table 1: WER results using RASTA-PLP features and MLPs trained in clean or heavy reverberation.

acceptable results. The advantage over the original product rule, as well as other related combination rules, is the ability to weight the streams either statically or dynamically based upon additional information.

3.1. Different MLPs with same preprocessing

Our first experiments test a multi-stream system where the individual probability streams are trained in different reverberant environments. RASTA-PLP[1] [5, 6] features with Δs and $\Delta\Delta$s were used as the feature extraction process for both streams. The features were normalized on a per-utterance basis to have a fixed mean and variance. The first two rows of table 1 show the word error rates (WER) for two streams where an MLP with 800 hidden units was trained in either the clean or heavy reverberation environment and then tested in these environments and a third light reverberation environment. The streams perform best in matched conditions. The third row shows the results when the probabilities from the two streams are combined using the log-average method with equal weights. In either of the matched cases (clean and heavy), this combination does not improve over the better of the individual streams, though fortunately the result is closer to the better stream than to the worse. In these cases, one stream is performing at its best while the other is at its worst. Of interest is the light reverberation test which is the unseen testing condition in this case. This score, marked with a "+", is significantly lower than either of the individual streams and is 21% better relative to the best stream. Since the combined stream implicitly has twice as many trained parameters as the individual streams, we also performed identical tests of the individual streams using a bigger MLPs with twice as many parameters. WER results are shown in the 4th and 5th rows of table 1. Increasing the number of parameters improves the word error rate in matched testing cases, though only significantly in the heavy reverberation case. Mismatched tests produce no significant difference in scores in these tests.

The last row of table 1 shows results where a single bigger MLP is trained with both clean and heavy data. The results are consistent with the log-average combination with only a significant difference in the heavy reverberation test case. The improved performance in the light reverberation test indicates that the MLP is able to do some interpolation between the two extreme training conditions. However, the matched test conditions experience some compromised performance just as the log-average method had. The simple posterior combination of heterogeneous streams appears more desirable than using a larger MLP trained in two conditions. The individual smaller MLP streams can be trained more quickly and in parallel. Additional streams can be added or removed without a complete retraining. More importantly, the simple posterior combination allows for weighting, where a stream can be de-selected based on additional knowledge. For example, in the matched cases, the mismatched stream could be given a weight of zero. Incorporating stream emphasis or de-emphasis would be less convenient when using a single trained MLP or when using an MLP as a stream merger.

[1] RelAtive SpecTrAl - Perceptual Linear Prediction

MLP Combination	WER (%) clean test	light test	heavy test
clean PLP alone	5.1	26.7	77.6
clean MSG alone	6.5	15.3	77.7
heavy PLP alone	39.2	31.4	35.4
heavy MSG alone	24.5	23.8	31.6
Homogeneous condition MLP tests			
clean PLP $\oplus$ MSG	4.3 +	14.9	70.4 +
heavy PLP $\oplus$ MSG	21.6 +	22.1 +	28.6 +
Heterogeneous condition MLP tests			
clean PLP $\oplus$ heavy MSG	5.9	14.7 +	43.8
log-average of all four	6.2	13.6 +	41.5

Table 2: WER Results from a frame level combination of PLP and MSG streams trained individually in clean and heavy reverberation. The "+" annotation marks where the combination is significantly better than the single streams. $\oplus$ signifies log-average combination.

3.2. Different MLPs with different preprocessing

Time and again it has been demonstrated that combinations of classifiers based upon features with different properties give rise to the best combination results, e.g. [4]. In addition to performing tests where the MLPs were trained in different environments, we conducted tests using different feature extraction processes. Our tests here use PLP cepstra [5] (without RASTA filtering) and MSG[2] lowpass and bandpass features [7]. PLP cepstral features together with Δs and $\Delta\Delta$s were normalized on a per-utterance basis to a fixed mean and variance. MSG features included 13 modulation features that were filtered with 8 Hz lowpass and 8-16 Hz bandpass IIR filters. The MSG features were also normalized but in an online manner. Four streams were trained individually on each set of features and with data from either of two acoustic conditions labeled "clean" and "heavy". These probability streams were tested individually and in combination using clean, light, and heavy reverberation conditions. Again, the "light" condition is an unseen test condition presented to the ASR system. Table 2 shows the word error rate (WER) from word recognition test results. Since different types of features sometimes require different values of decoding parameters to perform best, some parameter tuning was conducted using the CV set to achieve the lowest word error rate.

The first four rows of scores in table 2 give the WER of the individual streams. This row illustrates some of the performance differences between PLP and MSG. Overall, the MSG features perform better than PLP when the system was trained or tested on reverberated data. PLP on the other-hand performs better when training and testing on clean data. The next two rows of scores

[2] Modulation-filtered SpectroGram

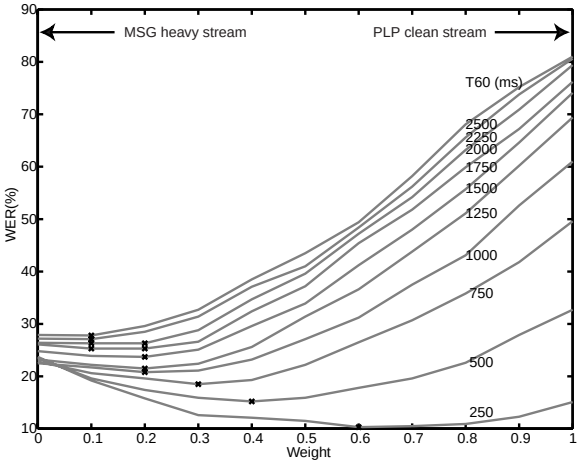

Figure 1: WER for a range of artificial reverberant conditions using a weighted log-average between PLP and MSG streams trained on clean and heavy reverberation data respectively. Tests performed on room impulses with DTR = -8 dB. "x" marks are placed on curve minima.

combine streams trained in like acoustic conditions using the log-average method and with tests in all three conditions. In most cases, the combination results in lower WER than the single streams, particularly in matched training and testing cases.

We subsequently performed tests using a heterogeneous environment combinations. First we used two streams where the PLP stream was trained with clean data and the MSG stream was trained with heavily reverberated data. This selection was chosen since PLP performed better on clean data and MSG performed better on reverberated data. The results shown in the second to the last row of table 2 reveal a similar pattern as with RASTA-PLP; specifically there is some compromise in performance in the matched cases coupled with an improvement in the unseen light reverberation test. The heterogeneous combination performs the same as the clean combination in the unseen light condition test. We note, however, that the improvement in the heterogeneous case over the best of the individual streams for the light test is more substantial. I.e. the best constituent stream is 23.8% for the heterogeneous case but 15.3% for the homogeneous case, yielding relative improvements of 38% and 2.6% respectively. Since combinations of PLP and MSG features demonstrates performance improvements in matched as well as mismatched testing conditions, we also tested using streams from both feature sets trained in both clean and heavy reverberation conditions. The last row of table 2 shows that the unseen light condition case improves further, though the scores are not significantly different from the two stream heterogeneous case in this test.

4. WEIGHTED COMBINATION

The previous experiments used a simple log-average of the probabilities with equal weighting. Since we use streams that were trained in separate environments, an equal weighting will not always be optimal; the speech test data may be a closer match to one of the streams. In the above cases, a weighting towards the clean stream in the clean test case or towards the heavy stream in the heavy test case would mitigate some of the compromised performance in the heterogeneous combination tests. We repeated tests using the two-streams heterogeneous case and a number of artificially constructed room impulse responses. The weight between the streams varied from 0 (all weight to the MSG-heavy stream) and 1 (all weight to the PLP-clean stream). Word recognition tests used the smaller CV set with fixed recognition parameters to speed up the evaluation. Figure 1 shows the WER results with the DTR held at -8 dB and with T60 varying from approximately 250 ms to 2500 ms. The existence of extrema in these curves is encouraging since it signifies that combinations of the streams in this fashion can produce lower WER than either stream alone. In the case of the smaller T60, corresponding to less reverberant quality in the speech data, the weighting should be more equal. For larger T60s at this DTR, the weighting should favor the heavy trained stream. A weighting knob to the ASR system, whether manual "rules of thumb" or automatically computed from statistical measures from the data, could be used to keep the system at peak performance.

5. FOUR-STREAM COMBINATION WITH UNSEEN ROOM IMPULSES

Final tests were conducted using the room impulse responses gathered from four microphones in a varechoic chamber with 100%, 43%, and 0% of the panels set "open" to an absorbent state. The streams were trained using clean data and heavy reverberation data and are independent environments from the training room impulses for these tests. The individual stream WER results are tabulated in table 3 with the WER of a log-average combination of all streams listed in the last column. The first 8 rows corresponding to the 100% and 43% open panels had lighter reverberation quality. From figure 1, the original equal weighting is adequate and all streams were given a 0.25 weight. The combination lowers the word error in all cases, by as much as 30% relative. The last four rows, corresponding to all panels in a reflective state, have a heavier reverberation quality. An equal weighting yielded results that were sometimes worse than the individual streams. From figure 1, a clean-stream weight of 0.2 would be considered more appropriate. These tests therefore assigned the 0.2 weight equally between the clean streams (0.1 each) and the 0.8 weight equally between the heavy streams (0.4 each). This weighting scheme lowered word error in all of these cases.

6. DISCUSSION

By and large, the environment in which the MLP probability estimator is trained is the overriding factor on the performance in different reverberant environments. Even with robust feature extraction routines, the difference in the feature distributions due to the room characteristics still varies wide enough to cause severe degradation in word recognition. For example, it seems unlikely that even given an arbitrarily large number of trained parameters, that the performance in heavy reverberation of a clean-trained system would approach the performance of the heavy-trained system. In fact it would be in danger of over-fitting the training data. Clean-trained systems tend to be only effective in approximately clean environments; the effects of more severe reverberant environments on the feature distributions is too great. Streams trained on some type of reverberation can sometimes have better results with other types of reverberation when compared to the performance of the clean-trained stream. The feature distributions may not deviate as far from it than from the clean and hence better probability estimates could be obtained. A possible partial solution, therefore, is to use a system trained with examples of what the distribution might look like in more than one environment.

Room impulse test				WER %				
Panels open(%)	Mic.	DTR (dB)	T60 (s)	PLP stream clean	PLP stream heavy	MSG stream clean	MSG stream heavy	log-average combination
100	1	1	0.3	12.0	42.3	9.6	24.4	8.6 +
100	2	1	0.3	10.6	43.0	9.2	23.5	7.7 +
100	3	-1	0.3	12.1	39.9	10.8	24.1	8.9 +
100	4	-1	0.3	11.7	41.6	10.1	24.7	9.8
43	1	1	0.5	22.2	30.8	13.9	23.7	11.0 +
43	2	-3	0.5	21.4	33.3	15.3	23.3	11.7 +
43	3	-2	0.5	22.3	31.1	17.6	23.6	11.9 +
43	4	-5	0.5	24.2	31.4	18.9	23.5	13.5 +
0	1	0.3	1	55.7	33.2	42.5	26.1	21.1 +
0	2	0.3	1	55.8	31.5	45.6	25.0	21.3 +
0	3	0.3	1	60.6	33.8	52.3	26.7	24.4 +
0	4	0.3	1	59.9	35.9	53.0	26.8	25.2 +

Table 3: Final tests using four PLP and MSG streams trained in clean and heavy reverberation. The combinations for the 100% and 43% open panels rows have equal weighting. For the 0% open panels, the clean streams were given a weight of 0.1 and the heavy, a weight of 0.4. The "+" annotations mark where the log-average combination produced WER that was significantly better than the single streams.

The most encouraging results occur when a two-stream system has streams that are trained in different environments and then presented with data from an unseen environment. Scores in the unseen environment are superior in the combined system than the singly trained systems. The unfortunate side effect is a penalty in either of the matched conditions. In these cases, the mismatched stream harms the combination more than helps, though the resulting combination still performs closer to the better stream. An extra input or measure that can intelligently switch off the worse stream appropriately would rectify problems in the matched cases. When such information is not available, however, the compromised performance in matched cases can still be a reasonable trade-off. Real deployed ASR systems will almost surely be presented with speech in an environment different from its training environment. Multiple streams trained in heterogeneous environments can broaden the range of graceful performance degradation.

7. CONCLUSION

A benefit of the multi-stream approach is that it can capitalize on the strengths of more than one approach for maintaining robustness to acoustic degradation. Advances in any of the components of the system can be readily integrated into the multi-stream system. The form used here employed multiple front-end acoustic modeling stages whose acoustic probability estimates were then merged for further processing by the word recognition decoder. Each of the front-end stages was trained to improve phone posterior estimation in a particular acoustic environment. Our results using clean and heavily reverberated training data show that this approach can aid ASR when presented with test data in an unseen reverberated environment. Addition of an appropriately set weighting knob can further help such a system operate at a reasonable performance level. With multiple front-end acoustic modeling streams trained in different conditions, the range of environments where the ASR system can maintain reasonable performance can increase.

8. ACKNOWLEDGMENTS

We would like to expressly thank Brian Kingsbury and Carlos Avendaño for making the room impulses available for our use. We would also like to thank Nelson Morgan and ICSI for supporting this work. This work was funded by NSF, grant number (NSF)-IRI-9712579.

9. REFERENCES

[1] J. A. Bilmes and K. Kirchhoff. Directed graphical models of classifier combinations: application to phone recognition. In *ICSLP*, Beijing, China, October 2000.

[2] H. Bourlard and N. Morgan. *Connectionist Speech Recognition- A Hybrid Approach*. Kluwer Academic Press, 1994.

[3] Center for Spoken Language Understanding, Department of Computer Science and Engineering, Oregon Graduate Institute. Numbers corpus, release 1.0, 1995.

[4] D. P. Ellis and J. A. Bilmes. Using mutual information to design feature combinations. In *ICSLP*, Beijing, China, October 2000.

[5] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.

[6] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[7] B. E. D. Kingsbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25(1-3):117–32, August 1998.

[8] J. Kittler, M. Hataf, R. Duin, and J. Matas. On combining classifiers. *IEEE Trasactions on Pattern Analysis and Machine Intelligence*, 3(20):226–39, 1998.

[9] T. Robinson and J. Christie. Time-first search for large vocabulary speech recognition. In *ICASSP*, Seattle, Washington, May 1998. IEEE.

[10] W. C. Ward, G. W. Elko, R. A. Kubli, and W. C. McDougald. The new varechoic chamber at AT&T Bell Labs. In *Proceedings of the Wallace Clement Sabine Centennial Symposium*, pages 343–6, Woodbury, NY, 1994. Acoustical Society of America.