RECOGNIZING REVERBERANT SPEECH WITH RASTA-PLP

Brian E. D. Kingsbury and Nelson Morgan

International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, CA 94704, USA
Dept. of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94704, USA
{bedk,morgan}@icsi.berkeley.edu

ABSTRACT

The performance of the PLP, log-RASTA-PLP, and J-RASTA-PLP front ends for recognition of highly reverberant speech is measured and compared with the performance of humans and the performance of an experimental RASTA-like front end on reverberant speech, and with the performance of a PLP-based recognizer trained on reverberant speech. While humans are able to reliably recognize the reverberant test set, achieving a 6.1% word error rate, the best RASTA-PLP-based recognizer has a word error rate of 68.7% on the same test set, and the PLP-based recognizer trained on reverberant speech has a 50.3% word error rate. Our experimental variant on RASTA processing provides a statistically significant improvement in performance on the reverberant speech, with a best word error rate of 64.1%.

1. INTRODUCTION

Robustness to reverberation in automatic speech recognition (ASR) systems is a problem of both practical and theoretical interest. One of the most promising aspects of ASR technology is the potential for hands-free interaction with machines. If this potential is to be fulfilled, however, the problem of reliably recognizing reverberant speech must be solved. Users will, quite reasonably, expect their systems to work equally well whether they are spoken to from across the room or from nearby. Reverberation is also interesting because it is a form of distortion quite distinct from both additive noise and spectral shaping. Unlike additive noise, reverberation creates interference that is correlated with the speech signal, and although reverberation and spectral shaping are both forms of linear convolutional distortion, spectral shaping is multiplicative in the short-time Fourier transform domain when a typical-length analysis window of around 10 ms is used, while reverberation is not. In a spectrographic display such as Figure 1, reverberation appears as a form of temporal smearing.

While humans are relatively tolerant of reverberation in speech, it appears that ASR systems are not. In [1], Sandhu and Ghitza reported that the phone error rate on TIMIT sentences increased from 27.1% on a clean test set to 81.3% on a reverberant test set for a recognizer that used a mel-cepstral front end. Using an auditory-based front end, the ensemble interval histogram (EIH), they found that the phone error rate increased from 36% on a clean test set to 82.7% on a reverberant test set. In that study, the reverberant test set was produced by processing the clean test set through a room reverberation simulator with a reverberation time of roughly 250–300 ms. In contrast, for humans listening to monaural presentations of Modified Rhyme Test (MRT) words in a carrier phrase, the error rate increased from 1.2% on a clean test set to 6.6% on a reverberant test set for word-initial consonants, and from 5.6% on a clean test set to 15.5% on a reverberant test set for word-final consonants [2]. In that study, the reverberant test set was produced by playing the clean test set in a room with an average reverberation time of 800 ms.

A RASTA-like algorithm for the dereverberation of speech was previously used to improve the performance of a speaker-dependent, isolated-word dynamic-time-warp recognizer on reverberant speech [3]. Our goal is not to enhance speech, but rather to extract reverberation-robust features for use in speaker-independent, continuous speech recognition systems. We are specifically interested in a class of generalized RASTA algorithms that use more moderate forms of automatic gain control than the current algorithms, that use methods other than autoregressive modeling for spectral smoothing and enhancement of spectral peaks, and that perform an analysis of temporal modulations in multiple modulation frequency bands. To provide a baseline reference for a study of these algorithms, we have tested PLP [4], log-RASTA-PLP, and J-RASTA-PLP [5] front ends, both singly and in combination, on a highly reverberant test set. We have included a simple example of the class of algorithms we are now studying in these tests, and the initial results are promising. We have also performed a simple human listening experiment on the same test set, under conditions that are closely matched to the machine recognition tests, to obtain an upper bound for performance. The following sections describe the human and machine recognition experiments, present the results of those experiments, and discuss their implications for performing robust ASR in reverberant environments.

2. EXPERIMENTAL CONDITIONS

2.1. Speech Material

Both the machine recognition and human recognition tests were performed using material from the Numbers93 corpus, a subset of the Numbers corpus [6] collected by the Center for Speech and Language Understanding at the Oregon Graduate Institute. The Numbers93 corpus is a collection of spontaneous utterances from many speakers, collected over the telephone and sampled at 8 kHz with a 16-bit A/D converter. The vocabulary is restricted to numbers and a few other words; a sample utterance from the corpus is "nine double oh one eight."

2.2. Generating Reverberant Speech

Reverberant speech for the experiments was generated by digitally convolving clean speech from Numbers93 with a hand-designed impulse response. The impulse response was

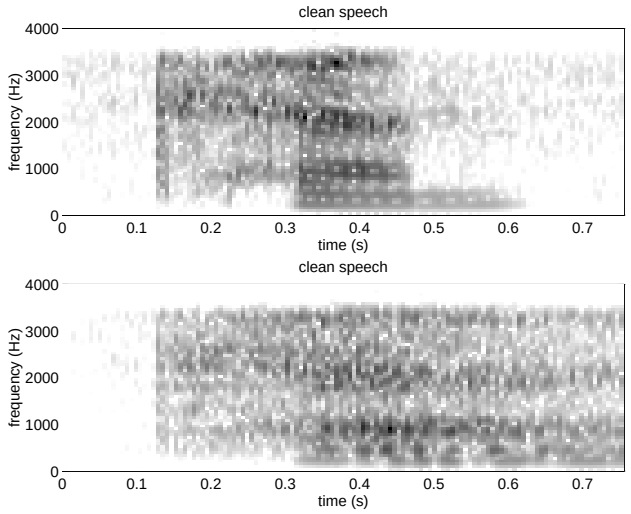

Figure 1. Spectrograms for clean and reverberant versions of the telephone-bandwidth utterance "ten."

designed to match the gross characteristics of a hallway about 6.1 m long, 2.4 m high, and 1.7 m wide, with concrete walls, floor, and ceiling. From a recording of speech collected in this hallway, reverberation times in different subbands were estimated. These estimates are summarized in Table 1. Next, Gaussian white noise was processed through an FIR filter bank to split it into subbands identical to those in which the reverberation times were measured, and each noise band was modulated with an exponential that matched the reverberation time estimated for that subband. The modulated noise bands were then added together to produce the reverberant tail of the impulse response. The sparse, early reflections were estimated using a simple time-domain point image expansion simulation [7]. The ratio of direct to reverberant sound was adjusted by ear to match the original recording conditions, in which the microphone was located approximately 2.5 m from the speaker. The ratio of direct to reverberant sound energy is -16 dB.

2.3. Machine Recognition Experiments

Machine recognition tests were run using a hybrid hidden Markov model/multilayer perceptron (HMM/MLP) recognizer [8] in which phone probabilities are estimated using an MLP and speech decoding is done with a Viterbi search. Four front ends were tested: PLP, log-RASTA-PLP, J-RASTA-PLP, and an experimental RASTA-like front end we call the modulation spectrogram. The modulation spectrogram is distinguished from other RASTA-PLP algorithms in a number of ways: the modulation spectrogram is computed using a filterbank for spectral analysis instead of a short-time Fourier transform, it uses off-line spectral normalization over an entire utterance instead of on-line adaptation, it performs an analysis of slow modulations in speech in the linear domain instead of the log or lin-log domain, and it uses global thresholding for enhancement of spectral peaks instead of autoregressive modeling. Further details on the modulation spectrogram are provided in [9].

The PLP, log-RASTA-PLP, and J-RASTA-PLP front ends used a 25 ms analysis window, an 80 Hz frame rate, and produced nine cepstral coefficients (including the energy term) per frame. The MLP phonetic probability estimators used with these front ends had 120 inputs (the cepstral coefficients, excluding the energy term, for the current frame, the previous seven frames, and the next seven frames), 512 hidden units, and 56 output units, for a total of 90,112 weights. The modulation spectrogram produced fifteen spectral coefficients per frame at an 80 Hz frame rate. The MLP phonetic probability estimator used with the modulation spectrogram features had 225 inputs (the spectral coefficients for the current frame, the previous seven frames, and the next seven frames), 320 hidden units, and 56 output units, for a total of 89,920 weights. A class bigram grammar language model was used during speech decoding. Each recognizer was trained on a set of 875 utterances. An iterative procedure, in which the speech was relabeled via forced alignment and a recognizer was trained using the new labels, was employed to ensure a good match between the features and word models used in recognition. Testing of the recognizers was carried out on clean and reverberant versions of a 657-utterance test set.

Freq. Band	Reverb. Time
0–250 Hz	3.1 s
250–500 Hz	2.6 s
500–1000 Hz	2.2 s
1000–2000 Hz	1.6 s
2000–4000 Hz	1.4 s

Table 1. Estimated reverberation times in different frequency bands for hallway.

Three different machine recognition experiments were run. First, to establish baseline results, recognizers using each of the four front ends were trained on a clean version of the training set, then tested on the clean and reverberant test sets. Second, recognizers using the PLP and modulation spectrogram features were trained on a reverberant version of the training set, then tested on the clean and reverberant test sets. Finally, recognition tests were run using frame-level phonetic probability estimates that were obtained by combining the estimates from MLPs trained on two different feature sets. The probability combination was accomplished by multiplying together the probability estimates from each MLP and normalizing by the prior probabilities. This technique is a version of a "mixture of experts" system, and works best when the MLPs used make independent errors. We hypothesize that when an MLP is unable to classify a given frame correctly, it may produce a relatively flat distribution of phone probabilities at its output. In this case, if the other MLP is able to correctly classify the frame, then the multiplication of its peaked output distribution by the other network's flat output distribution will not significantly change the probability estimates.

2.4. Human Recognition Experiments

For the human recognition tests, three subjects were asked to word-transcribe the same 657 reverberant utterances upon which the automatic recognition systems were tested. The subjects were provided with a list of the words present in Numbers93 because the automatic recognizers were provided with the same information. The order of the sentences was randomized to prevent any learning of speaker characteristics, and the subjects were allowed to listen to each utterance as many times as they wanted to minimize short-term memory effects. During the testing, the subjects were not given any feedback on their transcription accuracy. The utterances were produced by the 16-bit D/A converter in

experiment	feature set	condition	substitutions	deletions	insertions	error
baseline	PLP	clean	11.1%	3.2%	3.5%	17.8%
		reverb	35.0%	33.8%	2.7%	71.5%
	log-RASTA	clean	10.9%	3.0%	2.5%	16.4%
		reverb	41.0%	31.4%	2.0%	74.4%
	J-RASTA	clean	12.0%	3.2%	1.8%	16.9%
		reverb	46.0%	30.0%	2.9%	78.9%
	mod. spec.	clean	22.6%	6.6%	2.5%	31.7%
		reverb	42.2%	20.5%	3.4%	66.0%
train on reverb	PLP	clean	35.3%	4.2%	34.2%	73.3%
		reverb	29.2%	13.4%	7.6%	50.3%
	mod. spec.	clean	34.8%	6.8%	4.8%	46.4%
		reverb	30.9%	10.4%	3.2%	44.4%
combined probabilities	PLP & log-RASTA	clean	7.0%	2.6%	2.2%	11.9%
		reverb	36.4%	29.7%	2.6%	68.7%
	PLP & mod. spec.	clean	8.4%	2.2%	3.1%	13.6%
		reverb	39.0%	19.6%	5.5%	64.1%
humans		reverb	4.1%	1.4%	0.6%	6.1%

Table 2. Machine and human recognition results. All percentages are word error rates.

a SPARC-5 workstation at a sampling rate of 8 kHz, and were presented over headphones at a comfortable level in a quiet office. All subjects were native speakers of American English, had no known hearing impairments, and had considerable experience performing phonetic transcription of speech. Prior to the recognition testing, each subject was trained on ten other reverberant utterances from Numbers93 to familiarize them with the task. During training the subjects were given feedback on their transcription accuracy.

3. RESULTS

The results of the machine and human recognition tests are summarized in Table 2.

Among the baseline clean test results, the differences between the PLP, log-RASTA, and J-RASTA scores are not statistically significant. The score for the modulation spectrogram features is significantly worse. The differences between the scores on the baseline reverberant test are all statistically significant, with the modulation spectrogram features giving the best performance, followed by the PLP features.

The experiments in which the recognizer was trained on reverberant speech, then tested on clean and reverberant speech, illustrate the difficulty of the reverberant test set: even when the training and test conditions are matched, the recognizer word error rate is in the 44–50% range.

Combining phone probability estimates from the MLPs trained on the PLP and log-RASTA features, which is in some ways analogous to supplementing the PLP features with delta features, gives the best performance of any recognizer on the clean test, and also gives a statistically significant improvement in performance on the reverberant test. Combining phone probabilities from the MLPs trained on the PLP and modulation spectrogram features gives better performance on the clean test than any other recognizer except the combined PLP and log-RASTA recognizer, and gives the best performance on the reverberant test. The combined recognizers have twice as many parameters as the baseline recognizers, but their improved performance is not due to the larger number of weights: doubling the number of weights in the baseline recognizers does not yield a significant improvement in performance.

The human listeners had much less difficulty on this task than the automatic recognizers.[1] The human error rates are high for a word recognition task with a known vocabulary. Although we do not have measurements for human subjects listening to the clean test set, we note that the word error rate for humans listening to connected digit strings from the TI DIGITS corpus was 0.105% [10]. The error rate for humans on the reverberant test set is still lower than the best recognizer's error rate on the *clean* test set.

4. CONCLUSIONS

Recognition of highly reverberant speech is a difficult task. The best ASR system's score on the reverberant test set, with training on the clean test set, is an order of magnitude higher than the scores of human listeners. Furthermore, the score for the PLP-based recognizer that was trained then tested on reverberant speech is extremely high: 50%. This matched reverberant training and test result may be regarded as a lower bound on the performance of a PLP-based recognizer that uses a frame-oriented hybrid HMM/MLP architecture. Neither the log-RASTA nor the J-RASTA front end provide any improvement on their own on this reverberant test set, and while a combined PLP and log-RASTA recognizer does provide a significant improvement

[1] The human listeners' error rates were extremely consistent, as well as extremely low compared to the automatic recognizers. The word error rates on the listening test for each subject were as follows:

substitutions	108	103	91
deletions	25	35	40
insertions	13	15	14
total error	146	153	145

There were a total of 2426 words in the test.

in performance on the reverberant test set, it does not approach this lower bound.

The modulation spectrographic features, which are an experimental variant on RASTA, provide improved performance on the reverberant test set over the other front ends, and when the modulation spectrographic features are combined with PLP features, it is also possible to get good performance on the clean test set. We believe that the performance improvement in reverberation that is provided by the modulation spectrographic features arises from their robust representation of syllabic segments in the speech signal. This hypothesis is supported by the fact that most of the improvement in reverberation with the modulation spectrographic features comes from a reduction in the deletion rate. Also, we have observed that the modulation spectrographic features tend to highlight the high-energy portions of the speech signal associated with syllabic nuclei.

A variant on the modulation spectrographic features that does not perform any global thresholding [9] gives nearly the same performance on the clean speech as PLP, without any probability combination, and performance on the reverberant speech that is essentially the same as that provided by the modulation spectrographic features reported in this paper. A recognizer that uses these variant features, but is otherwise identical to the baseline modulation spectrographic recognizer described here has a word error rate of 19.2% on the clean test set and a word error rate of 66.9% on the reverberant test set. However, when this variant is combined with PLP, we do not observe the same performance improvement on clean and reverberant speech that we obtained here through combining.

It is possible that improvements in the front end may not, on their own, be sufficient to achieve reliable recognition of reverberant speech: in the matched reverberant training and test experiment with modulation spectrographic features, the word error rate is still 44%. Changes in the recognizer architecture, for example recognition based on syllables instead of phones, may be needed to cope with the long-time effects of room reverberation.

REFERENCES

[1] Sumeet Sandhu and Oded Ghitza. A comparative study of mel cepstra and EIH for phone classification under adverse conditions. In *ICASSP-95. The 1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 409–412. IEEE, 1995.

[2] Stanley A. Gelfand and Shlomo Silman. Effects of small room reverberation upon the recognition of some consonant features. *Journal of the Acoustical Society of America*, 66(1):22–29, July 1979.

[3] H. G. Hirsch. Automatic speech recognition in rooms. In J. L. Lacoume, A. Chehikian, N. Martin, and J. Malbos, editors, *Signal Processing IV: Theories and Applications. Proceedings of EUSIPCO-88. Fourth European Signal Processing Conference.*, volume 3, pages 1177–1180, Amsterdam, 1988. Elsevier Science Publishers B.V.

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

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

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

[7] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, 65(4):943–950, April 1979.

[8] Hervé Bourlard and Nelson Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[9] Steven Greenberg and Brian E. D. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. In *ICASSP-97. 1997 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1997.

[10] R. Gary Leonard. A database for speaker-independent digit recognition. In *ICASSP-84. 1984 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 42.11.1–42.11.4. IEEE, 1984.