THE EFFECTS OF BACKGROUND MUSIC ON SPEECH RECOGNITION ACCURACY

Bhiksha Raj, Vipul N. Parikh, and Richard M. Stern

Department of Electrical and Computer Engineering & School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213, USA

ABSTRACT

Recognition of broadcast data, such as TV and radio programs is a topic of great interest. One of the problems with such data is the frequent presence of background music that degrades the performance of speech recognition systems.

In this paper we examine the effects of different kinds of music on automatic speech recognition systems by comparing the effects of music with the relatively well-known effects of white noise on these systems. We also examine the extent to which compensation algorithms that have been successfully applied to noisy speech are also helpful in improving recognition accuracy for speech that is corrupted by music. It is hoped that these experimental comparisons will lead to a better understanding of how to compensate for the effects of background music.

1. INTRODUCTION

Speech recognition of broadcast news shows has received a great deal of recent attention (*e.g.* [1]). One attribute of broadcast speech is the presence of music in the background of large sections of data. While isolating these segments is a problem in itself, a far more difficult problem is frequently that of actually recognizing the speech content of these segments. For example, on the Marketplace data provided for the 1995 ARPA Hub 4 evaluations, the recognition accuracy obtained by the CMU SPHINX-II (semi-continuous HMM-based) speech recognition system obtained for studio-recorded speech with no background noise or music was 74.2%, while the recognition accuracy obtained for studio-recorded speech corrupted by music was only 64.7% [2]. Similar results were obtained by other participating sites as well (*e.g.* [3, 4].

One of the salient features of music as a corrupting signal is its non-stationarity. Consequently, compensation methods that have been successfully applied to speech corrupted by quasi-stationary noise may not necessarily improve speech recognition accuracy to the same extent when the corrupting signal is music.

In this paper we attempt to analyze the effects of music on the recognition accuracy obtained by speech recognition systems, and to isolate the impact of several different attributes of music as a corrupting signal that may make its effects different from those of stationary noise. To this end we progressively process white noise to resemble a particular segment of music in increasing degrees, and we evaluate the effect on recognition accuracy of each step of this process. We believe that better knowledge of how these individual attributes of musical sound contribute to the observed degradation in recognition accuracy may facilitate the development of algorithms to compensate for them.

In Section 2 we present baseline results for various kinds of music with and without noise compensation. In Section 3 we present a further analysis of one of the musical signals used in the experiments of Section 2. In Sections 4 we provide a brief discussion of how our experimental results may impact on appropriate compensation strategies for speech in the presence of background music. All wind and percussion music samples were obtained from the "Hollywood Edge" CD recorded by Tonal Images Inc. The market place music was obtained my manually editing all music segments from five shows of the 1995 ARPA Hub 4 development data. All experiments were performed using the 104-word speaker-independent continuous-speech alphanumeric Census database [6] using the CMU SPHINX-II semi-continuous HMM-based speech recognition system.

2. SPEECH RECOGNITION WITH BACKGROUND MUSIC

We evaluated the overall effects of three different kinds of music on speech recognition accuracy by corrupting high-quality digitized speech with the music and then recognizing the corrupted speech using a system that was trained using clean speech. The speech samples for each of these experiments were the testing utterances of the CMU Census database. The three sets of background music used for these experiments were (1) single notes of one of a randomly selected set of wind instruments, (2) single notes of one of a randomly selected set of percussion instruments, and (3) segments of music from test data in the 1995 ARPA Hub 4 evaluation, which consisted of episodes from the NPR program Marketplace. To establish a comparison, we also evaluated recognition accuracy using speech corrupted by white noise. Each test was conducted at several different SNRs, where SNR was defined to be the ratio of the overall energy of the speech to the overall energy of the corrupting music or noise. Finally, at each SNR, for each type of corrupting music, we also compared recognition accuracy obtained when the corrupted speech was processed by the CDCN algorithm [5, 6] an algorithm that jointly compensates for the effects of additive noise and linear filtering.

Results of these initial experiments are plotted in Figure 1. It can be seen that at the lower SNRs the impact of background music on recognition accuracy depends strongly on the type of music that is presented. The Hub 4 background music, which is the most complex and "natural", provided the greatest degradation and the wind instruments provided the smallest degradation in recognition accuracy. Nevertheless, none of the background music segments produced as much degradation as did white noise with the same SNR.

While the CDCN algorithm is quite effective in ameliorating the effects of degradation produced by additive white noise, the improvement in performance on the speech corrupted by music provided by CDCN processing is almost negligible. Interestingly, CDCN is more effective for the music from the Marketplace shows than for the simpler music types. CDCN is an algorithm

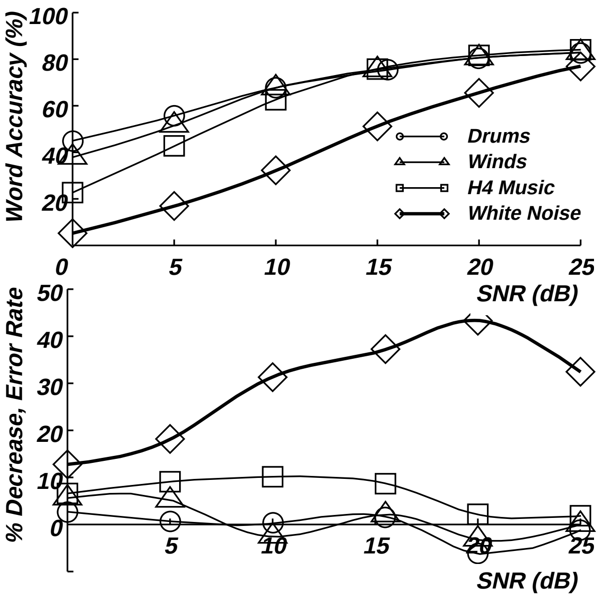

Figure 1. Upper panel: Recognition accuracy as a function of SNR for speech corrupted with various kinds of music. Lower panel: Percentage improvement in error rate provided by the CDCN algorithm.

that assumes a model for the degradation effected by a corrupting signal. Both data-driven compensation algorithms such as RATZ [6] and complete retraining of the HMMs using music-corrupted speech provided similar results, in that the improvement in recognition accuracy with music is not comparable to that obtained for white noise under identical conditions.

In recognition of the observation that the more complex music provides greater degradation of recognition accuracy, we progressively added instruments to the corrupting music in a second experiments and compared the resulting effects on recognition accuracy at several SNRs. These results are plotted in Figure 2. Though this is not a perfect simulation of music, it is reasonable to conclude from Figure 2 that as the corrupting music becomes more complex, recognition accuracy degrades.

3. ANALYSIS OF MUSIC

As we noted in Section 2, the effect of music on recognition accuracy approaches that of white noise as the music becomes more complex. Furthermore, compensation seems to be more effective as the music becomes more complex, and compensation is most effective for white noise. This leads us to believe that the more complex music is closer to white noise in some sense than the less complex ones.

In this section we attempt to identify various attributes of music that may make it different from white noise in terms of its effect on speech recognition accuracy. We attempt to isolate each of these features and process white noise to have each of them, in order to evaluate the features' effect on recognition accuracy.

While music differs from white noise in several ways, the following two differences are among the most obvious: (1) The average short-term power of stationary noise is, in principle, a constant independent of time. This, however, is not true for music. For example, Figure 3 depicts short-term power as a function of time for white noise, a drum note and a wind instrument. (2) The spectrum of music generally varies with frequency (and usually has strong harmonic structure), while the spectrum of white noise is constant by definition. Furthermore, the music spectrum is nonstationary and changes continually with time.

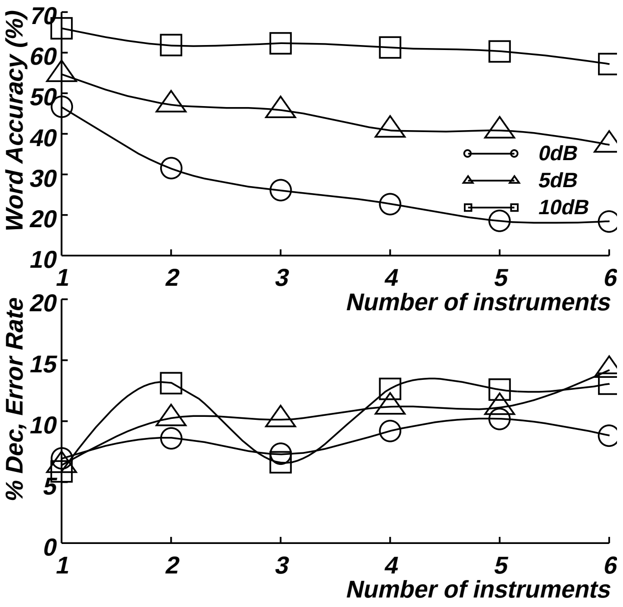

Figure 2. Upper panel: Effect of number of background instruments on recognition accuracy. Lower panel: Percentage improvement in error rate provided by the CDCN algorithm.

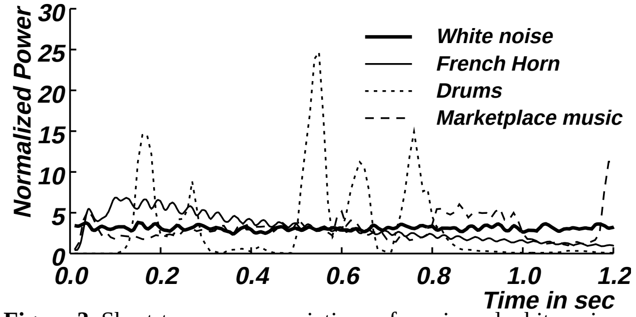

Figure 3. Short-term power variations of music and white noise with respect to time.

In the experiments described below we compare the effects of variation of short-term power and of spectral envelope on recognition accuracy. These experiments all make use of repetitions of a xylophone note as the music sample. While the xylophone was selected because its sound is both percussive and harmonic, similar results were observed for pilot experiments using a variety of other types of music.

3.1. EFFECTS OF POWER VARIATION

To evaluate the effect of the power variation of music on speech recognition accuracy we amplitude modulated a white noise signal so that it would produce the same power track as that of a xylophone note, on a frame-by-frame basis. The resulting recognition accuracy is shown in Figure 4. We observe that the application of the power track of music to white noise reduces the amount of degradation at all SNRs.

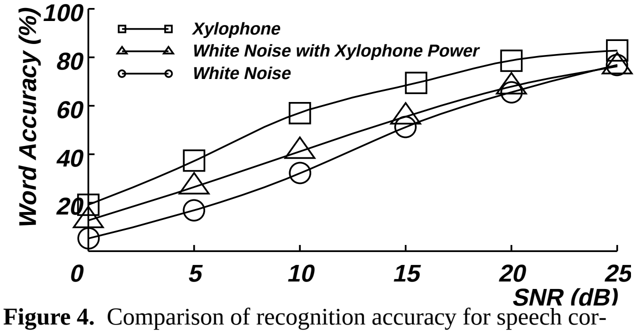

Figure 4. Comparison of recognition accuracy for speech corrupted by xylophone music with the power tracks of xylophone music and white noise.

To confirm the effect of power fluctuations we also performed the reverse experiment, measuring the recognition accuracy obtained for speech in the presence of background music that had been amplitude normalized to produce constant energy in each frame. This forces the power track of the music to be similar to that of the white noise. These results are plotted in Figure 5. As expected, we observe that normalizing the energy reduces recognition accuracy (This effect was much more pronounced when other music tracks were used as for the corrupting signal).

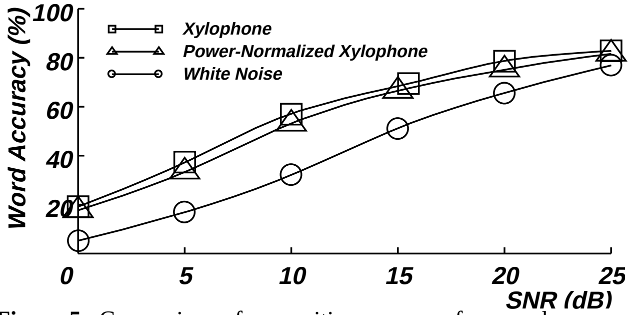

Figure 5. Comparison of recognition accuracy for speech corrupted by white noise with the power tracks of xylophone music and white noise.

We also observe that CDCN compensation provides a much greater improvement in recognition accuracy when the corrupting signal is power normalized, compared to the negligible effect provided by compensation for the non-normalized music. This appears to indicate that the effectiveness of compensation methods like CDCN is reduced when an additive corrupting signal exhibits fluctuations in power. (This is not unexpected, since these compensation procedures are based on an assumption that the background noise is quasi-stationary in nature.)

The preceding results all confirm that musical signals with large short-term power fluctuations degrade speech recognition accuracy less than stationary noise. One possible explanation for this observation is that the temporal regions of music lower instantaneous power result in regions of correspondingly higher SNR for the corrupted speech. These “islands” of higher SNR may be dominant in their effects over the other regions of lower SNR, providing a net improvement in recognition accuracy compared to the white noise case. In other words, corrupting signals with regions of low power may degrade recognition accuracy less than signals with more constant power at the same overall SNR.

3.2. EFFECTS OF SPECTRAL CHARACTERISTICS

Musical noise and white noise also differ in their spectral variation. We could break this down further into the average spectrum of music which could be considerably different from that of white noise, and the actual variation in this spectrum with time. It would be reasonable to extrapolate from the observations in the previous section that power variations along the frequency axis may have the same effect as power variations in time: frequency bands of high SNR could compensate for the effects of frequency bands of low SNR.

We filtered white noise to have the same average spectrum as a xylophone note, and we used this colored noise as a corrupting signal to evaluate the effect of average spectrum of music on recognition accuracy. The power in any frequency band of the music signal also varies as a function of time, and it is not obvious that band-level fluctuations in power should have the same effect as fluctuations in the overall power of the signal on speech recognition accuracy. To evaluate the salience of band-level fluctuations in music power, white noise was processed using LPC analysis to have the same spectral pattern as a xylophone note. This results in a signal that has a frequency-smoothed version of the spectrum of the music.

Results obtained using these signals as maskers are described in Figure 6, and summarized for the SNR of 10 dB in Table 1. The colored noise with the average spectrum of the xylophone note is observed to cause lower degradation than white noise, while the noise that follows the xylophone spectrum is observed to have almost the same effect as the xylophone note itself.

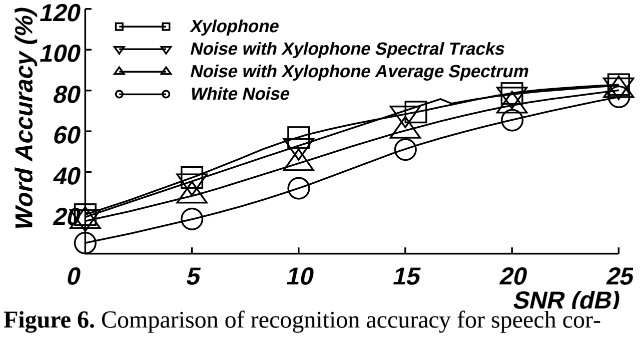

Figure 6. Comparison of recognition accuracy for speech corrupted by xylophone music, plus noise with the time-varying spectral tracks of the xylophone, noise with the time-averaged spectrum of the xylophone, and white noise.

Corrupting signal Type	No CDCN	CDCN
White noise	32.3%	53.5%
White noise + xyl. pwr. track	39.3%	39.5%
White noise + xyl. avg. spectrum	44.4%	73.3%
White noise +xyl. spec. and pwr. track	52.9%	53.5%
Xylophone	57.0%	57.4%

Table 1. Summary of recognition accuracies at SNR 10 dB, with and without the use of the CDCN algorithm.

As the corrupting noise approaches music, the degradation of baseline accuracy is seen to diminish at each step, and no single attribute appears to be substantially more important than any of the others. However, power fluctuations in general seems to have a significant effect on the effectiveness of CDCN compensation, as the results in Table 1 indicate. The use of CDCN results in improved recognition accuracy only for cases in which the corrupting signal is stationary. Compensation with RATZ[6] resulted in better accuracy than CDCN at lower SNRs; however, the improvement over CDCN was negligible and did not compare with the improvements obtained under similar circumstances when the corrupting signal was white noise.

4. DISCUSSION

Our experiments seem to indicate that any deviation in the nature of a corrupting signal from white noise, whether it be spectral or temporal, results in improved speech recognition accuracy. Regions of low power in the corrupting signal, both in frequency and in time, seem to compensate for regions of high power to provide better uncompensated recognition accuracy compared to the white noise case.

Nevertheless, compensation methods such as CDCN or RATZ tend to fail if the power or spectrum of the corrupting signal vary with time. These algorithms assume stationarity either explicitly or implicitly, which is clearly not valid in the cases of music or white noise that is processed to include time-varying amplitude or spectral fluctuations.

The combination of these observations seems to suggest that systems that recognize speech that has been corrupted with non-stationary noise sources such as music should give greater weight to regions of high SNR more than regions of low SNR in their recognition strategy. Approaches that ignore the acoustic evidence from regions of very low SNR completely (*e.g.* [7, 8, 9]) may work well. The problem, of course, lies with actually locating these regions.

On the other hand, noise compensation methods should be able to account for the temporal patterns of the corrupting signal. A-two dimensional HMM (*e.g.* [10]) is one such possibility. The problem here would be the need for prior knowledge of the HMM of the music, or the ability to learn its parameters from the test data itself. Another compensation possibility would be methods that whiten the corrupting signals before compensating for them.

ACKNOWLEDGEMENTS

This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

REFERENCES

1. Rudnicky, A., “Hub 4: Business Broadcast News”, Proc. ARPA Speech Recognition Workshop, pages 8-11, February, 1996
2. Jain, U., Siegler, M. A., Doh, S.-J., Gouvea, E., Moreno, P. J., Raj, B., and Stern, R. M., “Recognition of continuous broadcast news with multiple unknown speakers and environments”, Proc. ARPA Speech Recognition Workshop, pages 61-66, February, 1996.
3. Steven Wegmann et. al. (1996) “Marketplace recognition using Dragon’s continuous speech recognition system”, Proceedings of the ARPA speech recognition workshop, pages 67-71, February, 1996.
4. P.S.Gopalakrishnan et.al. (1996), “Suppressing background music from music-corrupted data of the ARPA Hub-4 task”, Proceedings of the ARPA speech recognition workshop, pages 81-84, February, 1996.
5. Acero, A., *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Boston, MA, 1993.
6. Moreno, P. J., Raj, B., Gouvea, E.B., and Stern, R. M., “Multivariate-gaussian-based Cepstral Normalization for Robust Speech Recognition, *ICASSP- 95*, pages 137-140, May, 1995.
7. M. P. Cooke, A. Morris, P. D. Green (1996). “Recognizing occluded speech”, Proc. ESC ETRW on the auditory basis of speech perception, 1996.
8. Hermansky, H., Tibrewala, S., and Pavel, M., “Towards ASR on Partially Corrupted Speech”, Proc. ICSLP-96, pages 544-547, October, 1996
9. Bourlard, H. and Dupont, S., “A New ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands”, Proc. ICSLP-96, pages 629-632, October, 1996
10. A. P. Varga, A. P., and Moore, R. K., “Hidden Markov Model Decomposition of Speech and Noise”, *ICASSP- 90*, pages 845-848, April, 1990.