



Time-Compressed Speech Perception with Speech and Noise Maskers

Douglas S. Brungart and Nandini Iyer

Air Force Research Laboratory, Wright Patterson AFB, Ohio

douglas.brungart@wpafb.af.mil, nandini.iyer@wpafb.af.mil

Abstract

Many researchers have shown that speech signals can be time compressed (TC) by a factor of two or more without a significant loss in intelligibility. However, most previous studies with TC speech have been conducted either in quiet or, in a very small number of cases, with noise maskers. In this experiment, we examine the effect that TC has on the perception of a speech signal in the presence of a speech or noise masker. The results show that normal speech can be accelerated a modest amount (20-30%) without increased susceptibility to masking, but that higher TC ratios can lead to dramatically worse performance in the presence of an interfering sound. The results also indicate that time-expansion can, in some cases, lead to improved performance when a listener is attending to the quieter of two talkers in an auditory mixture. These results suggest that there are some important practical limitations on how TC should be used to enhance communications efficiency in auditory speech displays.

Index Terms: time compressed speech, multitalker perception

1. Introduction

Our ability to communicate effectively in crowded environments like restaurants or cocktail parties is highly dependent on our ability to extract a single speech signal of interest from a background comprised of many simultaneous interfering voices. Most normal hearing listeners are able to routinely accomplish this impressive signal processing task in complex acoustic environments that would cause complete failure for even the most advanced automatic speech recognition systems. Furthermore, there is evidence that human listeners are able to unconsciously scan the content of the interfering voices and switch their attention to a new target voice when an especially pertinent piece of information, like an unexpected utterance containing their own name, arises in the acoustic background.

Because multitalker listening is such an important component of effective human communication, a great deal of research has been focused on identifying the acoustic features that listeners use to segregate competing voices. Of course, one very important component of multitalker listening is the overall target-to-masker (TMR) ratio of the target speech. In all cases, speech perception accuracy improves when the TMR increases above 0 dB. However, when the target speech signal is masked by a single masking voice, performance also tends to improve when the TMR *decreases* below 0 dB [1]. This phenomenon has been attributed to the ability of listeners to selectively focus their attention on the quieter of two talkers in a multitalker stimulus. Other factors that are known to influence multitalker speech perception are the relative similarity of the target and masking voices [2], and the spatial separation between the target and masking voices [3].

One aspect of multitalker speech perception that has thus

far received relatively little attention is the role that speaking rate plays in the segregation of competing voices. Intuitively, it is clear that speaking rate has a profound impact on speech perception, particularly for hearing impaired listeners and for those attempting to understand a non-native language. However, speaking rate is a very difficult parameter to examine with unaltered speech recordings, because normal talkers are only able to adjust their speaking rate over a very limited range without introducing artifacts into the spoken utterance. Thus, most studies that have examined the impact of speaking rate on speech perception, including some dating back to the late 1950's, have used some form of electronic speech compression to generate their stimuli [4].

Studies that have used TC speech fall into two broad categories: the first group of studies addresses the usefulness of TC speech signals as an efficient speech-based display. As such, the studies have examined the impact that a change in speaking rate has on intelligibility for a normal-hearing listener in a quiet environment. These studies suggest that listeners can accurately comprehend clean speech that is accelerated to roughly double its normal speed. For example, Versfeld and Dreschler [5] found that listeners could identify speech with near 100% accuracy when it was accelerated from a normal rate of 3-4 syllables per second to a TC rate of roughly 8 syllables per second. Others have estimated that listeners can understand connected speech spoken at a maximum rate of roughly 200 words per minute, versus a normal speaking rate of roughly 120 words per minute [6, 7].

The second group of studies used TC speech signals as a way to study the auditory system. To this end, these studies have evaluated the effect of various types of distortions on speech perception for normal and hearing-impaired listeners [8, 9].

Only a very small number of studies have directly or indirectly examined TC speech perception in the presence of a masker. The results of these studies suggest TC speech may be more sensitive to interference from a noise masker than normal speech. For example, Bornstein [10] examined the perception of TC speech (compressed stimulus duration decreased by 40% of the original stimulus duration) in quiet and in the presence of a 12-talker babble, and found that word intelligibility declined to a greater extent with TC speech than with normal speech.

To this point, we are unaware of any studies that have systematically examined the effect that TC has on a listener's ability to segregate a target speech signal from an acoustic background comprised of competing speech maskers, rather than noise or babble maskers. Nor is there any data to indicate what effect TC might have in situations where there is some ambiguity about the identities of the target and masking voices (e.g. where the listener is forced to use the content of the competing phrases to determine which competing voice is the target talker). In this paper, we describe an experiment that examined the effect of TC on two-talker speech segregation with stimuli

derived from the Coordinate Response Measure [1], a call-sign based speech perception test that requires listeners to listen to both talkers in the stimulus for the phrase containing the target call sign (“Baron”) and respond with the color-number combination contained in that target phrase. As a baseline condition, performance was also measured in the presence of a speech-spectrum-shaped noise. The results are discussed in terms of their implications for the use of TC as a means to develop enhanced multitalker speech display systems.

2. Methods

2.1. Stimuli

The stimuli consisted of phrases from the Coordinate Response Measure (CRM) corpus [1]. The corpus consists of eight talkers (4 male, 4 female) each saying phrases of the form Ready [call sign], go to [color] [number] now. Eight possible call signs, four possible colors and eight possible numbers result in 256 phrases per talker and 2048 total phrases in the corpus. The PSOLA algorithm, as implemented in the publicly available PRAAT speech processing software package [11], was used to uniformly stretch or compress the original phrases from the corpus. In this experiment, PRAAT was used to time compress or expand the original CRM phrases, which varied in length from 1.57 s to 2.28 s (mean 1.8 s), to one of seven predetermined phrase lengths: 0.4, 0.64, 1, 1.25, 1.5, 2 and 3 seconds. A control condition was also tested where the CRM phrases were presented at their original durations. In this condition, the signals were processed through the PSOLA algorithm, but the original duration of the phrase was maintained to capture any artifacts introduced by the PSOLA processing.

2.2. Procedure

The listeners heard the signals presented diotically over Beyerdynamic DT 990 Pro headphones at comfortable listening levels while seated in a quiet room. On each trial, a target phrase denoted by the call sign Baron was presented along with either a speech or noise masker. Listeners had to respond to the color-number combination in the target phrase by a mouse click on a response grid displayed on a computer monitor. The response grid comprised of eight columns of numbers arranged in four rows each depicting the colors in the CRM corpus. Feedback was provided after every response, and overall percentage correct in a block was graphically depicted at the end of each block. Signal generation and presentation and the recording of responses were controlled using MATLAB software.

When the masker was speech, another CRM phrase that differed in call sign, color and number combination from the target phrase was presented along with the target phrase. Both phrases were either stretched or compressed to the same predetermined phrase length on each trial. In order to control the similarity of the target and masking voices, the masking phrase was equally likely to be spoken by the same talker as the target phrase, by a different talker of the same sex as the target talker, or by a talker different in sex than the target talker. When the masker was noise, a speech masker chosen in the same way described above was multiplied by a broadband Gaussian noise in the frequency domain and then inverse fast Fourier transformed to yield an unintelligible random-phase masker with the same spectral content as the corresponding speech masker. The overall RMS levels of these signals were scaled to produce seven different target-to-masker ratios (TMRs) ranging from -12 dB to 12 dB in 4-dB steps. A target-alone control condition was

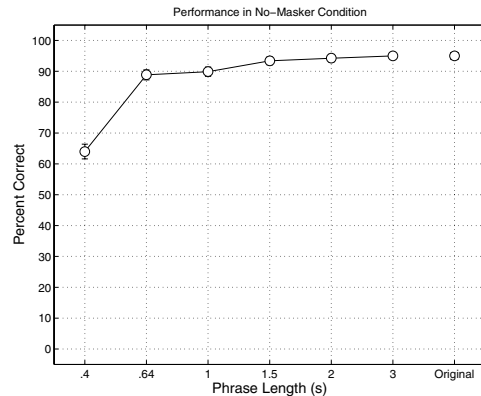


Figure 1: Correct color and number identification performance as a function of phrase length in trials with no masking stimulus. The mean errors were calculated separately for each listener, and then averaged together to yield the overall values in the figure. Each data point represents a minimum of 411 trials, and the error bars represent ± 1 standard error around each data point.

also included to evaluate intelligibility in quiet. Within a block of trials, the type of masker (i.e., speech or noise) remained fixed and TMR as well as the phrase length were randomized. Data were collected on a total of nine normal-hearing listeners (4 male and 5 female) with ages ranging from 21 to 56 years.

3. Results

3.1. Effect of TC with No Masking Signal

Figure 1 shows how performance in the experiment varied as a function of phrase length in the control condition with no masker. The data are plotted in terms of the percentage of trials where the listener correctly identified both the color and the number in the stimulus. From these data, it is apparent that TC had little or no impact on the intelligibility of the CRM stimuli in quiet until the phrase length was reduced to less than 0.64 seconds long. Since the original phrases were, on average, approximately 1.8 seconds long, this represents a compression ratio of approximately 3 to 1. When the phrase length was reduced to 0.4 s (a compression ratio of 4.5 to 1), performance degraded to 60% correct responses, but was still significantly better than chance (1/32 or roughly 3%). These results are consistent with other studies that have shown that listeners can tolerate quite high TC ratios when they are listening to speech stimuli in quiet.

3.2. Effects of TC as a Function of TMR

Figure 2 shows how overall performance varied across all the conditions tested in the experiment. The left panel shows performance with the CRM speech masker, while the right panel shows performance with the random-phase noise masker. Within each panel, each different curve shows performance as a function of TMR for a different phrase length (as indicated by the legend). Although the data are complicated, there are a few important trends that are immediately apparent from this figure.

- When the target phrase was presented at its original length (right-pointing triangles in Figure 2), the perfor-

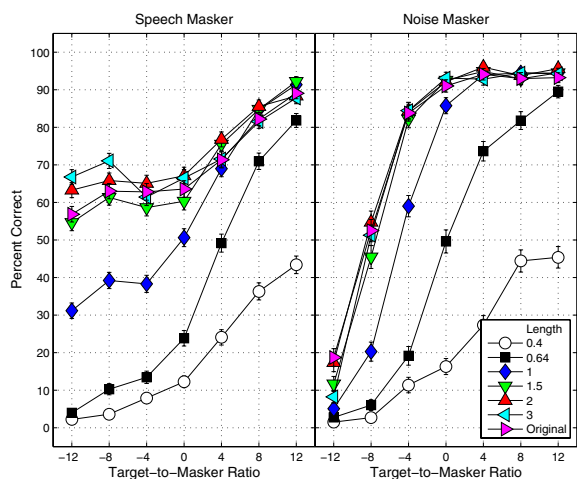


Figure 2: Performance as a function of TMR for each phrase length tested in the experiment. The left panel shows data with the speech masker, and the right panel shows data with the noise masker. Each data point represents a minimum of 304 trials, and the error bars represent ± 1 standard error around each data point.

performance curves were quite different for the speech and noise maskers. With the noise maskers (right panel), performance was very good at TMR values down to 0 dB, and began to decrease rapidly as the TMR was reduced below 0 dB. In the speech masking condition (left panel), performance dropped off gradually as the TMR was reduced from 12 dB to 0 dB, and then tended to plateau at TMR values less than 0 dB. The relatively high level of performance that occurs at negative TMR values for a speech masker is believed to be related to the use of a strategy where the listener selectively focuses attention on the “quieter” talker in the stimulus [1].

- Increasing the phrase length beyond 1.5 s generally had little or no impact on performance with either masker at any TMR. This can be seen by the tight grouping of the top four curves in each panel of Figure 2. The only exception was a slight improvement in performance for time-expanded speech at low TMR values in the speech masking condition. This is discussed in more detail in the next section.
- Time-compressing the stimuli to a length of one second (a compression ratio of roughly 1.8 to 1) significantly impacted performance only when the TMR was 0 dB or less. At TMR values below 0 dB, TC to a length of 1 s reduced performance in the noise condition by the equivalent of roughly a 4 dB decrease in TMR. In the speech condition, it reduced performance even more dramatically, especially at very low TMR values. This suggests that TC may interfere with the listener’s ability to selectively attend to the quieter talker in a two-talker stimulus.
- Time-compressing the stimuli to a length of 0.64 s (a compression ratio of roughly 3 to 1) caused some degradation in performance even when the TMR was +12 dB. When the TMR value was reduced below +12 dB, changing the phrase length from 1 s to 0.64 s produced a re-

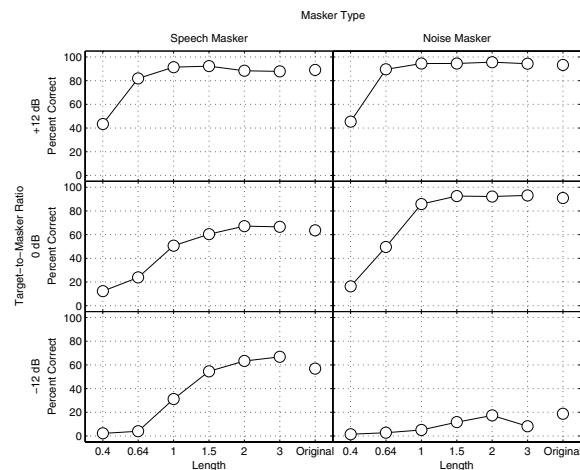


Figure 3: Performance as a function of phrase length at three TMRs. The left panel shows data with the speech masker. The right panel shows data with the noise masker. The error bars represent ± 1 standard error around each data point.

duction in performance equivalent to a 4-10 dB decrease in the TMR.

- Time-compressing the stimuli to a length of 0.4 s reduced performance at all TMR values. Notably, performance in the 0.4 s phrase-length condition was virtually identical in the speech and noise masking conditions of the experiment.

Thus, to summarize, it seems that TC ratios of up to roughly 2 to 1 can be tolerated in conditions where there is no masker or a very weak masker, but that even compression ratios smaller than 2 to 1 can lead to a substantial decrease in performance when the target speech signal is masked by an interfering speech or noise signal. Furthermore, it seems that these modest compression ratios can also severely compromise a listener’s ability to selectively attend to the quieter talker in a two-talker stimulus. Compression ratios greater than 3 to 1 seem to significantly impair performance at all but the highest signal-to-noise ratios. Thus, it seems that these very high compression ratios should only be used in situations where the listener is expected to hear the resulting speech signals in a quiet environment.

3.3. Effects of TC as a function of TMR and Masker Type

Figure 3 shows performance in the experiment as a function of phrase length for each of three different TMRs. The top row of the figure shows performance at a relatively high TMR value (+12 dB). In this case, as in Figure 1, we see that TC has very little impact on performance until the compression ratio exceeds roughly 3 to 1. We also see that time expansion (e.g. the 2 s and 3 s phrase lengths) has no positive benefit at high TMR values.

The second row of the figure shows performance at an intermediate TMR value of 0 dB. In this case, it is clear that TC starts to have some impact on performance when the phrase length is reduced to 1 s or less (a compression ratio of 1.8 to 1).

The last row of the figure shows performance at a very low TMR value of -12 dB. In this condition, performance is very poor in the noise masking condition at all phrase lengths, but reasonably good ($> \approx 50\%$ correct responses) at phrase lengths greater than 1.5 s in the speech masking condition. In this

case, it is very clear that time expansion from the original 1.8 s phrase length to 3 seconds produced some improvement in performance (from roughly 50% correct responses to roughly 60% correct responses). Thus, it seems that time expansion *can* provide some benefit in situations where a listener is trying to attend to the quieter of two competing phrases. Notably, no such benefit is seen in the noise condition. In fact, performance in the 3 s phrase length condition is roughly 10% *worse* than in the original condition with the noise masker. This suggests that time expansion can sometimes provide some performance benefit for speech segregation tasks, but that it provides little or no benefit for listeners who are attempting to understand a speech signal masked by noise.

4. Discussion and Conclusions

In environments where TC speech is listened to in the presence of a noise masker, there is a clear performance trade-off between the TC ratio and the target-to-masker ratio. Speech can be compressed a modest amount (roughly 20-30%) without increasing its sensitivity to noise, but higher compression ratios tend to lead to dramatically worse intelligibility performance in the presence of a masker¹. This effect is likely related to the loss of redundant acoustic information in TC speech. Some of the earliest experiments with TC speech were motivated by the observation that listeners could understand periodically “interrupted speech even when the speech signal was presented at relatively low duty-cycles (i.e., 10 ms of speech followed by 20 ms of silence). This result suggested that listeners did not require all the information present in a speech signal to understand it, and that some of this information could be removed (by TC) without impairing intelligibility. However, it is likely that this redundant information *is* needed to understand speech in the presence of noise, because some of the acoustic information will inevitably be lost due to spectral overlap between the target speech and the masker. Thus, it is perhaps not surprising that interfering noise would impair performance with TC speech.

When the masking sound is a competing speech utterance, the overall effects of TC are similar to those that occur with a noise masker. TC seems to degrade performance at roughly the same compression factor (1.8) for both speech and noise. However, the net effect of TC appears to be greater at low TMR ratios for a speech masker, because listeners generally start from a higher level of performance than they do for a noise masker. However, one aspect of performance that does appear to be different for speech and noise maskers is the effect of time expansion. There is no evidence that listeners ever benefit from the time expansion of a speech signal presented in noise, but at low TMR values there is some evidence that listeners might benefit slightly from time expansion of a speech signal presented in the presence of a speech masker.

In terms of practical applications, these results strongly suggest that there are some real limits in terms of how time compressed speech might be used to enhance speech display systems. There is little doubt that TC can improve efficiency for a listener attending to a long passage of pre-recorded speech in a quiet listening environment. This approach could even work in cases where the listener is monitoring multiple channels of pre-recorded speech, so long as the speech channels are interleaved in such a way that they are presented sequentially,

¹It is perhaps a lucky accident that this is approximately the same amount of compression that can be achieved when talkers intentionally try to ‘speak quickly.

rather than simultaneously. However, many real-world applications are likely to require a listener to attend to TC pre-recorded speech while at the same time monitoring other live communications channels, and in these cases it is likely that listeners will begin to sacrifice some performance when the speech is TC by more than a modest amount. Performance will also likely be negatively impacted for users who listen to TC speech in noisy environments. Future audio display designers who are interested in using TC speech to enhance operator efficiency will need to seriously consider these potential real-world limitations of this very promising human interface technology.

5. Acknowledgements

Portions of this work were supported by AFOSR LRIR 01-HE-01-COR.

6. References

- [1] D.S. Brungart, “Informational and energetic masking effects in the perception of two simultaneous talkers,” *Journal of the Acoustical Society of America*, vol. 109, pp. 1101–1109, 2001.
- [2] C.J. Darwin and R.W. Hukin, “Effectiveness of spatial cues, prosody, and talker characteristics in selective attention,” *Journal of the Acoustical Society of America*, vol. 107, pp. 970–977, 2000.
- [3] A. Bronkhorst, “The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,” *Acustica*, vol. 86, pp. 117–128, 2000.
- [4] G. Fairbanks and F. Kodman, “Word intelligibility as a function of time compression,” *Journal of the Acoustical Society of America*, vol. 29, pp. 636–644, 1957.
- [5] N. J. Versfeld and W. A. Dreschler, “The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners,” *Journal of the Acoustical Society of America*, vol. 111, pp. 401–408, 2001.
- [6] L. Junor, “Time compressed speech: A resource alternative for nonvisual reading,” *Applis*, vol. 5, no. 1, pp. 23–27, 1992.
- [7] B. Arons, “Techniques, perception, and applications of time-compressed speech,” .
- [8] P.G. Lacroix and J.D. Harris, “Multiplicative effects on sentence comprehension for combined acoustic distortions,” *Journal of Speech and Hearing Research*, vol. 22, pp. 259–269, 1979.
- [9] S. Gordon-Salant and P.J. Fitzgibbons, “Effects of stimulus and noise rate variability on speech perception by younger and older adults,” *Journal of the Acoustical Society of America*, vol. 115(14), pp. 1808–1817, 2004.
- [10] S.P. Bornstein, “Time compression and release from masking in adults and children,” *Journal of the American Academy of Audiology*, vol. 5, pp. 89–98, 1994.
- [11] P. Boersma and D. Weenink, “Praat, a system for doing phonetics by computer, version 3.4,” *Institute of Phonetic Sciences, University of Amsterdam*, vol. 134, pp. 1–182, 1996.