Techniques, Perception, and Applications of Time-Compressed Speech

Barry Arons
Speech Research Group, MIT Media Lab
20 Ames Street, E15-353, Cambridge MA 02139
+1 617-253-2245
barons@media-lab.mit.edu

1 Abstract

There are a variety of techniques for time-compressing speech that have been developed over the last four decades. This paper consists of a review of the literature on methods for time-compressing speech, including related perceptual studies of intelligibility and comprehension.

2 Motivation and Applications

The primary motivation for time-compressed speech is for reducing the time needed for a user to listen to a message—to increase the communication capacity of the ear. A secondary motivation is that of data reduction—to save storage space and transmission bandwidth for speech messages.

Time-compressed speech can be used in a variety of application areas including teaching, aids to the disabled, and human-computer interfaces. Studies have indicated that listening to teaching materials twice that have been speeded up by a factor of two is more effective than listening to them once at normal speed [Sti69]. Time-compressed speech has been used to speed up message presentation in voice mail systems [Hej90, Max80], and in aids for the blind. Speech can be slowed for learning languages, or for the hearing impaired. Time compression techniques have also been used in speech recognition systems to time normalize input utterances to a standard length [Mal79].

While the utility of time compressing recordings is generally recognized, surprisingly, its use has not become pervasive. Rippey performed an informal study on users of a time-compression tape player installed in a university library. Virtually all the comments were positive, and the librarians reported that the speech compressor was the most popular piece of equipment in the library [Rip75].

The lack of commercial acceptance of time-compressed speech is partly because of the cost of compression devices and the quality of the reproduced speech, but is also attributable to the lack of user control. Traditionally, recordings were reproduced at fixed compression ratios where "... the rate of listening is completely paced by the recording and is not controllable by the listener. Consequently, the listener cannot scan or skip sections of the recording in the same manner as visually scanning printed text, nor can the listener slow down difficult-to-understand portions of the recording" [Por78].

Powerful computer workstations with speech input/output capabilities make high quality time-compressed speech readily available. It is now practical to integrate speech time-compression techniques into interactive voice applications, and the software infrastructure of workstations, portable, and hand-held computers to provide user interfaces for high-speed listening.

3 Considerations

There are three variables that can be studied in compressed speech [Duk74a]:

1. The type of speech material to be compressed (content, language, background noise, etc.).
2. The process of compression (algorithm, mono or stereo presentation, etc.).
3. The listener (prior training, intelligence, listening task, etc.).

Other related factors come into play in the context of integrating speech into computer workstations or hand-held computers:

1. Is the material familiar or self-authored, or is it unfamiliar to the listener?
2. Does the recorded material consist of many short items, or large unsegmented chunks of speech?
3. Is the user listening for maximum comprehension, or quickly skimming?

4 A Note on Compression Figures

There are several ways to express the amount of compression produced by the techniques described in this document. The most common figure in the literature is the compression percentage[1]. A compression of 50% corresponds to a factor of 2 increase in speed (requiring half

[1] An attempt has been made to present all numbers quoted from the literature in this format.

the time to play). A compression of 20% corresponds to a factor of 5 increase in speed. These numbers are most easily thought of as a percent reduction in time or data.

5 General Time-Compression Techniques

Time-compressed speech is also referred to as accelerated, compressed, time-scale modified, sped-up, rate-converted, or time-altered speech[2]. There are a variety of techniques for changing the playback speed of speech—a survey of these methods are described briefly in the following sections. Note that these techniques are primarily concerned with reproducing the entire recording, not scanning portions of the signal. Most of these methods also work for slowing speech down, but this is not of primary interest. Much of the research summarized here was performed between the mid 1950's and the mid 1970's, often in the context of accelerated teaching techniques, or aids for the blind.

5.1 Speaking Rapidly

The normal English speaking rate is between 130–200 words per minute (wpm). When speaking fast, a talker unintentionally changes relative attributes of his speech such as pause durations, consonant-vowel duration, etc. Talkers can only compress their speech to about 70% because of physiological limitations[3] [BM76].

5.2 Speed Changing

Speed changing is analogous to playing a tape recorder at a faster (or slower) speed. This method can be replicated digitally by changing the sampling rate during the playback of a sound. These techniques are undesirable since they produce a frequency shift proportional to the change in playback speed, causing a decrease in intelligibility.

5.3 Speech Synthesis

With purely synthetic speech it is possible to generate speech at a variety of word rates. Current text-to-speech synthesizers can produce speech at rates up to 550 wpm. This is typically done by selectively reducing the phoneme and silence durations. This technique is powerful, particularly in aids for the disabled, but is not relevant to recorded speech.

5.4 Vocoding

Vocoders that extract pitch and voicing information can be used to time-compress speech. Most vocoding efforts, however, have focused on bandwidth reduction rather than on naturalness and high speech quality. The phase vocoder, described in section 7.2, is an exception.

5.5 Silence Removal

A variety of techniques can be used to find silences (pauses) in speech and remove them. The resulting speech is "natural, but many people find it exhausting to listen to because the speaker never pauses for breath" [Neu78]. The simplest methods involve the use of energy or average magnitude measurements combined with time thresholds; other metrics include zero-crossing rate measurements, LPC parameters, etc. A variety of speech/silence detection techniques are reviewed in detail in [Aro92].

Maxemchuk [Max80] used 62.5ms frames of speech corresponding to disk blocks (512 bytes of 8kHz, 8-bit μ-law data). For computational efficiency, only a pseudo-random sample of 32 out of every 512 values were looked at to determine low energy portions of the signal. Several successive frames had to be above or below a threshold in order for a silence or speech determination to be made.

TASI (Time Assigned Speech Interpolation) is used to approximately double the capacity of existing transoceanic telephone cables [MS62]. Talkers are assigned to a specific channel while they are speaking; the channel is then freed during silence intervals. During busy hours, a talker will be assigned to a different channel about every other "talkspurt". The TASI speech detector is necessarily a real-time device, and must be sensitive enough to prevent clipping of the first syllable. However, if it is too sensitive, the detector will trigger on noise and the system will operate inefficiently. The turn-on time for the TASI speech detector is 5ms, while the release time is 240ms. The newer DSI (Digital Speech Interpolation) technique is similar, but works entirely in the digital domain. Note that Maxemchuk's system was primarily concerned with reducing the time a listener needed to hear a message and minimizing storage requirements. DSI/TASI are concerned with conserving network bandwidth.

More sophisticated energy and time heuristics ([LRRW81, RS75], summarized in [O'S87]) are used in endpoint detection for isolated word recognition—to ensure that words are not inadvertently clipped. The algorithms for such techniques are more complex than those mentioned above, and such fine-grained accuracy is probably not necessarily for compressed speech or speech scanning.

6 Time Domain Techniques

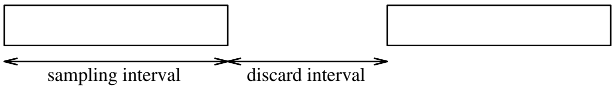
6.1 Sampling

The basis of much of the research in time-compressed speech was originated in 1950 by Miller and Licklider with experiments that demonstrated the temporal redundancy of speech. The motivation for this work was to increase channel capacity by switching speech on and off at

[2] "Time-scale modified" is often used in the signal processing literature, "time-compressed" or "accelerated" is often used in the psychology literature.

[3] According to the Guinness Book of World Records, John Moschitta has been clocked speaking at a rate of 586 wpm.

regular intervals so the channel could be used for another transmission (see figures 1 and 2B). It was established that if interruptions were made at frequent intervals, large portions of a message could be deleted without affecting intelligibility [ML50].



sampling interval	I_s
discard interval	I_d
time compression ratio	$R_c = I_d/(I_d + I_s)$

Figure 1: Sampling terminology [FK57]

Other researchers concluded that listening time could be saved by abutting the interrupted speech segments. This was first done by Garvey who manually spliced audio tape segments together [Gar53a, Gar53b], then by Fairbanks with a modified tape recorder with four rotating pickup heads[4] [FEJ54]. The bulk of literature involving the intelligibility and comprehension of time-compressed speech is based on such electromechanical tape recorders.

In the Fairbanks, or sampling, technique, segments of the speech signal are alternatively discarded and retained (figure 2C). This has traditionally been done isochronously—at constant sampling intervals without regard to the contents of the signal. Implementing such an algorithm on a general purpose processor is straightforward.

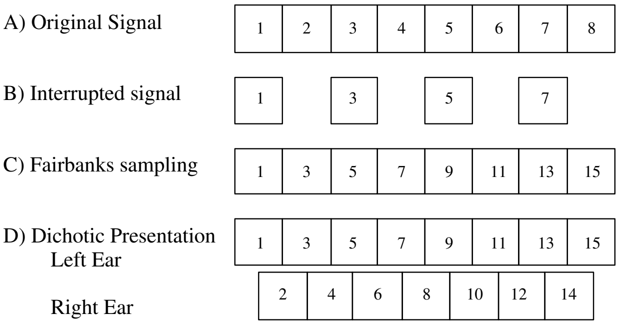


Figure 2: Sampling techniques

Word intelligibility decreases if I_d is too large or too small. Portnoff [Por81] notes that the duration of each sampling interval should be at least as long as one pitch period (e.g., $>$ 15ms), but should also be shorter than the length of a phoneme. Although computationally simple, such time-domain techniques introduce discontinuities at the interval boundaries that are perceived as "burbling" distortion and general signal degradation.

It has been noted that some form of windowing function or digital smoothing at the junctions of the abutted segments will improve the audio quality. The "braided-speech" method continually blended adjacent segments with linear fades, rather than abutting segments [Que74].

Lee describes two digital electronic implementations of the sampling technique [Lee72], and discusses the problems of discontinuities when segments are simply abutted together.

6.2 Sampling with Dichotic Presentation

One interesting variant of the sampling method (figure 2D) is achieved by playing the standard sampled signal to one ear and the "discarded" material to the other ear [5] ([Sco67] summarized in [Orr71]). Under this dichotic[6] condition, intelligibility and comprehension increase. Most subjects also prefer this technique to a diotic presentation of a conventionally sampled signal. Listeners initially reported a switching of attention between ears, but they quickly adjusted to this unusual sensation. Note that for compression ratios up to 50%, the two signals to the ears contain common information. For compressions greater than 50% some information is necessarily lost.

6.3 Selective Sampling

The basic sampling technique periodically removes pieces of the speech waveform without regard to whether it contains any redundant speech information. David and McDonald demonstrated a bandwidth reduction technique in 1956 that selectively removed (redundant) pitch periods from speech signals [DM56]. Scott applied the same ideas to time compression, setting the sampling and discard intervals to be synchronous with the pitch periods of the speech. Discontinuities in the time compressed signal were reduced, and intelligibility increased [SG72]. Neuburg developed a similar technique in which intervals equal to the pitch period were discarded (but not synchronous with the pitch pulses). Finding the pitch pulses is hard, yet estimating the pitch period is much easier, even in noisy speech [Neu78].

Since frequency-domain properties are expensive to compute, it has been suggested that easy-to-extract time-domain features can be used to segment speech into transitional and sustained segments. For example, simple amplitude and zero crossing measurements for 10ms frames can be used to group adjacent frames for similarity—redundant frames can then be selectively removed [Que74]. Toong [Too74] selectively deleted 50–90% of vowels, up to 50% of consonants and fricatives, and up to 100% of silence. However, he found that complete elimination of silences was undesirable (see also section 9.4).

[4] Several commercial products were based on this design, including one called the "Whirling Dervish."

[5] Usually with a delay of half of the discard interval.

[6] Dichotic refers to presenting different signals to each ear—note that headphones must be worn. Diotic is presenting the same signal to each ear, monotic is the presentation of a signal to only one ear.

"The most popular refinement of the Fairbanks technique is pitch-synchronous implementation (Scott; Huggins; Toong). Specifically, for portions of speech that are voiced, the sections of speech that are repeated or discarded correspond to pitch periods. Although this scheme produces more intelligible speech than the basic asynchronous pitch-independent method, errors in pitch marking and voiced-unvoiced decisions introduce objectionable artifacts (Toong)... Perhaps the most successful variant of the Fairbanks method is that recently proposed by Neuburg. This method uses a crude pitch detector, followed by an algorithm that repeats or discards sections of the speech equal in length to the average pitch period then smooths together the edges of the sections that are retained. Because the method is not pitch synchronous, and, therefore, does not require pitch marking, it is more robust than pitch-synchronous implementations, yet much higher quality than pitch-independent methods." [Por78]

6.4 Synchronized Overlap Add Method

The synchronized overlap add method (SOLA) first described by Roucos and Wilgus [RW85] has recently become popular in computer-based systems. It is a fast non-iterative optimization of a fourier-based algorithm described in [GL84]. "Of all time scale modification methods proposed, SOLA appears to be the simplest computationally, and therefore most appropriate for real-time applications" [WRW89]. Conceptually, the SOLA method consists of shifting the beginning of a new speech segment over the end of the preceding segment to find the point of highest cross-correlation. Once this point is found, the frames are overlapped and averaged together, as in the sampling method. This technique provides a locally optimal match between successive frames[7]; combining the frames in this manner tends to preserve the time-dependent pitch, magnitude, and phase of a signal. The shifts do not accumulate since the target position of a window is independent of any previous shifts [Hej90]. The SOLA method is simple and effective as it does not require pitch extraction, frequency-domain calculations, phase unwrapping, and is non-iterative [ME86]. The SOLA technique can be considered a type of selective sampling that effectively removes redundant pitch periods.

A windowing function can be used with this technique to smooth between segments, producing significantly less artifacts than traditional sampling techniques. Makhoul used both linear and raised cosine functions for averaging windows, and found the simpler linear function sufficient [ME86]. The SOLA algorithm is robust in the presence of correlated or uncorrelated noise, and can improve the signal to noise ratio of noisy speech [WW88, WRW89].

[7] The technique does not attempt to provide global optimality.

Several improvements to the SOLA method have been suggested that offer improved computational efficiency, or increased robustness in compression/decompression applications [ME86, WW88, WRW89, Har90, Hej90]. Hejna, in particular, provides a detailed description of SOLA, including an analysis of the interactions of various parameters used in the algorithm.

7 Frequency Domain Techniques

In addition to the frequency domain methods outlined in this section, there are a variety of other frequency-based techniques that can be used for time compressing speech (e.g., [MQ86, QM86]).

7.1 Harmonic Compression

Harmonic compression involves the use of a fine-tuned (typically analog) filter bank. The energy outputs of the filters are used to drive filters at half the frequency of the original. A tape of the output of this system is then played on a tape recorder at twice normal speed. The compression ratio of this frequency domain technique was fixed, and was being developed before the time when it was practical for digital computers to be used for time-compression.

Malah describes time-domain harmonic scaling which requires pitch estimation, is pitch synchronous, and can only accommodate certain compression ratios [Mal79, Lim83].

7.2 Phase Vocoding

A vocoder that maintains phase [Dol86] can be used for time-compression. A phase vocoder can be interpreted as a filterbank and thus is similar to the harmonic compressor. A phase vocoder is, however, significantly more complex because calculations are done in the frequency domain, and the phase of the original signal must be reconstructed.

Portnoff [Por81] developed a system for time-scale modification of speech based on short-time Fourier analysis. His system provided high quality compression of up to 33% while retaining the natural quality and speaker-dependent features of the speech. The resulting signals were free from artifacts such as glitches, burbles, and reverberations typically found in time-domain methods of compression such as sampling.

Phase vocoding techniques are more accurate than time domain techniques, but are an order of magnitude more computationally complex because Fourier analysis is required. The phase vocoder is particularly good at slowing speech down to hear features that cannot be heard at normal speed—such features are typically lost using time domain techniques. Dolson says "a number of time-domain procedures... can be employed at substantially less computational expense. But from a standpoint of fidelity (i.e., the relative absence of objectionable artifacts), the phase vocoder is by far the most desirable."

8 Combined Compression Techniques

The time-compression techniques described above can be mixed and matched in a variety of ways. Such combined methods can provide a variety of signal characteristics and a range of compression ratios.

8.1 Silence Removal and Sampling

Maxemchuk [Max80] found that eliminating every other non-silent block (1/16th second) produced "extremely choppy and virtually unintelligible playback." Eliminating intervals with less energy than the short-term average (and no more than one in a row), produced distorted but intelligible speech. This technique produced compressions of 33 to 50 percent. Maxemchuk says that this technique ". . . has the characteristic that those words which the speaker considered to be most important and spoke louder were virtually undistorted, whereas those words that were spoken softly are shortened. After a few seconds of listening to to this type of speech, listeners appear to be able to infer the distored words and obtain the meaning of the message." He believes such a technique would be "useful for users of a message system to scan a large number of messages and determine which they wish to listen to more carefully or for users of a dictation system to scan a long document to determine the areas they wish to edit."

Silence compression and sampling can be combined in several ways. Silences can first be removed from a signal that is then sampled. Alternatively, the output of a silence detector can be used to set boundaries for sampling, producing a selective sampling technique. Note that using silences to find discard intervals eliminates the need for a windowing function to smooth (de-glitch) the sound at the boundaries of the sampled intervals.

8.2 Silence Removal and SOLA

On the surface it would appear that removing silences and time-compressing speech using a technique such as the overlap-add method would be linearly independent, and could thus be performed in either order. In practice there are some minor differences, because the SOLA algorithm makes assumptions about the properties of the speech signal. The Speech Research Group has informally found a slight improvement in speech quality by applying the SOLA algorithm before removing silences. Note that timing parameters must be modified under these conditions. For example with speech compressed 50%, the silence removal timing thresholds must also be cut in half.

This combined technique is effective, and can produce a fast and dense speech stream. Note that silence periods can be selectively retained or shortened, rather than simply removed.

8.3 Dichotic SOLA Presentation

A sampled signal compressed by 50% can be presented dichotically so that exactly half the signal is presented to one ear, while the remainder of the signal is presented to the other ear. Generating such a lossless dichotic presentation is difficult with the SOLA method because the segments of speech are shifted relative to one another to find the point of maximum similarity. However, by choosing two starting points in the speech data carefully (based on the parameters used in the SOLA algorithm), it is possible to maximize the difference between the signals presented to the two ears. We have informally found this technique to be effective since it combines the high quality sounds produced with the SOLA algorithm with the binaural effect of the dichotic presentation.

9 Perception of Time-Compressed Speech

There has been a significant amount of perceptual work performed in the areas of intelligibility and comprehension of time-compressed speech. Much of this research has been summarized in [BM76], [FS69], and [Fou71].

9.1 Intelligibility vs. Comprehension

"Intelligibility" usually refers to the ability to identify isolated words. Depending on the type of experiment, such words may either be selected from a closed set, or written down (or shadowed) by the subject from an open-ended set. "Comprehension" refers to the understanding of the content of the material. This is usually tested by asking questions about a passage of recorded material.

In general, intelligibility is more resistant to degradation as a function of time-compression than is comprehension [Ger74]. Early studies showed that single well-learned phonetically balanced words could remain intelligible with a 10–15% compression (10 times normal speed), while connected speech remains comprehensible to a 50% compression (twice normal speed).

> "If speech, when accelerated, remains comprehensible the savings in listening time should be an important consideration in situations in which extensive reliance is placed on aural communication. However, current data suggest that although individual words and short phrases may remain intelligible after considerable compression by the right method, when these words are combined to form meaningful sequences that exceed the immediate memory span for heard words, as in a listening selection, comprehension begins to deteriorate at a much lower compression." [Fou71]

9.2 Limits of Compression

There are some practical limitations on the maximum amount that a speech signal can be compressed. Portnoff notes that arbitrarily high compression ratios are not physically reasonable. He considers, for example, a voiced phoneme containing four pitch periods. Greater than 25% compression reduces this phoneme to less than one pitch period, destroying its periodic character. Thus, he expects high compression ratios to produce speech with a rough quality and low intelligibility [Por81].

The "dichotic advantage" (section 6.2) is maintained for compression ratios of up to 33%. For discard intervals between 40–70ms, dichotic intelligibility was consistently higher than diotic intelligibility [GW77]. A dichotic discard interval of 40–50ms was found to have the highest intelligibility (40ms was described as the "optimum interval" in another study [Ger74]. Earlier studies suggest that a shorter interval of 18–25ms may be better for *diotic* speech [BM76]).

Gerber showed that 50% compression presented diotically was significantly better than 25% compression presented dichotically, even though the information quantity of the presentations was the same. These and other data provide conclusive evidence that 25% compression is too fast for the information to be processed by the auditory system. The loss of intelligibility, however, is not due to the loss of information because of the compression process [Ger74].

Foulke [FS69] reported that comprehension declines slowly up to a word rate of 275wpm, but more rapidly beyond that point. The decline in comprehension was not attributable to intelligibility alone, but was related to a processing overload of short-term memory. Recent experiments with French have shown that intelligibility and comprehension do not significantly decay until a high rate (300wpm) is reached [RSLM88].

Note that in much of the literature the limiting factor that is often cited is word rate, not compression ratios. The compression required to boost the speech rate to 275 words per minute is both talker and context dependent (e.g., read speech is typically faster than spontaneous speech).

Foulke and Sticht permitted sighted college students to select a preferred degree of time-compression for speech spoken at an original rate of 175wpm. The mean preferred compression was 82% corresponding to a word rate of 212wpm. For blind subjects it was observed that 64–75% time-compression and word rates of 236–275 words per minute were preferred. These data suggest that blind subjects will trade increased effort in listening to speech for a greater information rate and time savings [ZDS68].

Comprehension of interrupted speech (as in [ML50]) was good, probably because the temporal duration of the original speech signal was preserved, providing ample time for subjects to attempt to process each word [HLLB86]. Compression necessitates that each portion of speech be perceived in less time than normal. However, each unit of speech is presented in a less redundant context, so that more time per unit is required. Based on the large body of work in compressed speech, Heiman suggests that 50% compression removes virtually all redundant information. With greater than 50% compression, critical non-redundant information is also lost. They conclude that the compression ratio rather than word rate is the crucial parameter, because greater than 50% compression presents too little of the signal in too little time for a sufficient number of words to be accurately perceived. They believe that the 275 wpm rate is of little significance, but that compression and its underlying temporal interruptions decrease word intelligibility that results in decreased comprehension.

9.3 Training Effects

As with other cognitive activities, such as listening to synthetic speech, exposure to time-compressed speech increases both intelligibility and comprehension. There is a novelty in listening to time-compressed speech for the first time that is quickly overcome with experience.

Even naive listeners can tolerate compressions of up to 50%, and with 8–10 hours of training, substantially higher speeds are possible [OFW65]. Orr hypothesizes that "the review of previously presented material could be more efficiently accomplished by means of compressed speech; the entire lecture, complete with the instructor's intonation and emphasis might be re-presented at high speed as a review." Voor found that practice increased comprehension of rapid speech, and that adaptation time was short (minutes rather than hours) [VM65].

Beasley reports on an informal basis that following a 30 minute or so exposure to compressed speech, listeners become uncomfortable if they are forced to return to the normal rate of presentation [BM76]. He also reports on a controlled experiment extending over a six week period that found subjects' listening rate preference shifted to faster rates after exposure to compressed speech.

9.4 The Importance of Silence

> "Just as pauses are critical for the speaker in facilitating fluent and complex speech, so are they crucial for the listener in enabling him to understand and keep pace with the utterance." [Rei80]

> ". . . the debilitating effects of compressed speech are due as much to depriving listers of ordinarily available processing time, as to degradation of the speech signal itself." [WN80]

It may not be desirable to completely remove silences, as they often provide important semantic and syntactic cues. With normal prosody, intelligibility was higher for periodic segmentation (inserting silences after every eighth word[8]) than for syntactic segmentation (inserting silences after major clause and sentence boundaries) [WLS84]. Wingfield says that "time restoration, especially at high compression ratios, will facilitate intelligibility primarily to the extent that these presumed processing intervals coincide with the linguistic structure of the speech materials."

In another experiment, subjects were allowed to stop time-compressed recordings at any point, and were instructed to repeat what they had heard [WN80]. It was found that the average reduction in selected segment duration was almost exactly proportional to the increase in the speech rate. For example, the mean segment duration for the normal speech was 3s, while the chosen segment duration of speech compressed 60% was 1.7s. Wingfield found that "while time and/or capacity must clearly exist as limiting factors to a theoretical maximum segment size which could be held [in short-term memory] for analysis, speech content as defined by syntactic structure, is a better predictor of subjects' segmentation intervals than either elapsed time or simple number of words per segment. This latter finding is robust, with the listeners' relative use of the [syntactic] boundaries remaining virtually unaffected by increasing speech rate."

In the perception of normal speech, it has been found that pauses exerted a considerable effect on the speed and accuracy with which sentences were recalled, particularly under conditions of cognitive complexity [Rei80]. Pauses, however, are only useful when they occur between clauses within sentences—pauses within clauses are disrupting. When a 330ms pause was inserted ungrammatically, response time for a particular task was increased by 2s. Pauses suggest the boundaries of material to be analyzed, and provide vital cognitive processing time.

Maxemchuk found that eliminating silent intervals decreased playback time of recorded speech with compression ratios of 50 to 75 percent depending on the talker and material. In his system a 1/8 second pause is inserted whenever a pause greater or equal to 1 second occurred in a message. This appeared to be sufficient to prevent different ideas or sentences in the recorded document from running together. This type of rate increase does not affect the intelligibility of individual words within the active speech regions [Max80].

Studies of pauses in speech also consider the duration of the "non-pause" or "speech unit". In one study of spontaneous speech, the mean speech unit was 2.3 seconds. Minimum pause durations typically considered in the literature range from 50–800ms, with the majority in the 250–500ms region. As the minimum pause duration increases, the mean speech unit length increases (e.g, for pauses of 200, 400, 600, and 800ms, the corresponding speech unit lengths were 1.15, 1.79, 2.50, and 3.52s respectively). In another study, it was found that inter-phrase pauses were longer and occurred less frequently than intra-phrase pauses (data from several articles summarized in [Agn74]).

Hesitation pauses are not under the conscious control of the talker, and average 200–250ms. Juncture pauses are under talker control, and average 500–1000ms. Several studies have shown that breath stops in oral reading are about 400ms. In a study of the durational aspects of speech, it was found that the silence and speech unit durations were longer for spontaneous speech than for read speech, and that the overall word rate was slower. The largest changes occurred in the durations of the silence intervals. The greater number of long silence intervals were assumed to reflect the tendency for speakers to hesitate more during spontaneous speech than during oral reading [Min74].

Lass states that juncture pauses are important for comprehension, so they cannot be eliminated or reduced without interfering with comprehension [LL77]. Theories about memory suggest large-capacity rapid-decay sensory storage followed by limited capacity perceptual memory. Studies have shown that increasing silence intervals between words increases recall accuracy. Aaronson suggests that for a fixed amount of compression, it may be optimal to delete more from the words than from the intervals between the words. She states that "English is so redundant that much of the word can be eliminated without decreasing intelligibility, but the interword intervals are needed for perceptual processing" [AMS71].

10 Conclusions

This paper reviews of a variety of techniques for time-compressing speech, as well as related perceptual limits of intelligibility and comprehension. The SOLA method is currently favored for real-time applications, however, a digital version of the Fairbanks sampling method can easily be implemented and produces fair speech quality with little computation.

Time-compressed speech has recently begun showing up in voice applications and computer interfaces that use speech [WSB92]. Allowing the user to interactively change the speed at which speech is presented is important in getting over the "time bottleneck" often associated with voice interfaces. The techniques described in this paper can thus aid in user acceptance of voice applications.

[8] The silences were long (3s) in the context of the time-compression goals described in this document.

A Note on Important References

Though dated, the most readily accessible, and most often cited, reference is [FS69]. Another broad and more recent summary is [BM76]. An extensive anthology and bibliography [Duk74b] that contains copies and extracts of many earlier works is still in print.

Acknowledgements

Lisa Stifelman and Eric Hulteen provided comments on a draft of this paper.
This work was sponsored by Apple Computer, Inc.

References

[Agn74] J. G. Agnello. Review of the literature on the studies of pauses. In S. Duker, editor, *Time-Compressed Speech*, pages 566–572. Scarecrow, 1974.

[AMS71] D. Aaronson, N. Markowitz, and H. Shapiro. Perception and immediate recall of normal and "compressed" auditory sequences. *Perception and Psychophysics*, 9(4):338–344, 1971.

[Aro92] B. Arons. A review of adaptive speech detection, April 1992. Speech Research Group Memo.

[BM76] D. S. Beasley and J. E. Maki. Time- and frequency-altered speech. In N. J. Lass, editor, *Contemporary Issues in Experimental Phonetics*, chapter 12, pages 419–458. Academic Press, 1976.

[DM56] E. E. David and H. S. McDonald. Note on pitch-synchronous processing of speech. *Journal of the Acoustic Society of America*, 28(7):1261–1266, 1956.

[Dol86] M. Dolson. The phase vocoder: A tutorial. *Computer Music Journal*, 10(4):14–27, 1986.

[Duk74a] S. Duker. Summary of research on time-compressed speech. In S. Duker, editor, *Time-Compressed Speech*, pages 501–508. Scarecrow, 1974.

[Duk74b] S. Duker. *Time-Compressed Speech: an Anthology and Bibliography in Three Volumes*. Scarecrow, Metuchen, N.J., 1974.

[FEJ54] G. Fairbanks, W. L. Everitt, and R. P. Jaeger. Method for time or frequency compression-expansion of speech. *Transaction of the Institute of Radio Engineers, Professional Group on Audio*, AU-2:7–12, 1954. Reprinted in G. Fairbanks. *Experimental Phonetics: Selected Articles*, University of Illinois Press, 1966.

[FK57] G. Fairbanks and F. Kodman. Word intelligibility as a function of time compression. *Journal of the Acoustic Society of America*, 29:636–641, 1957. Reprinted in G. Fairbanks. *Experimental Phonetics: Selected Articles*, University of Illinois Press, 1966.

[Fou71] E. Foulke. The perception of time compressed speech. In P. M. Kjeldergaard, D. L. Horton, and J. J. Jenkins, editors, *Perception of Language*, chapter 4, pages 79–107. Charles E. Merrill Publishing Company, 1971.

[FS69] W. Foulke and T. G. Sticht. Review of research on the intelligibility and comprehension of accelerated speech. *Psychological Bulletin*, 72:50–62, 1969.

[Gar53a] W. D. Garvey. The intelligibility of abbreviated speech patterns. *Quarterly Journal of Speech*, 39:296–306, 1953. Reprinted in J. S. Lim *Speech Enhancement*, Prentice-Hall, Inc., 1983.

[Gar53b] W. D. Garvey. The intelligibility of speeded speech. *Journal of Experimental Psychology*, 45:102–108, 1953.

[Ger74] S. E. Gerber. Limits of speech time compression. In S. Duker, editor, *Time-Compressed Speech*, pages 456–465. Scarecrow, 1974.

[GL84] D. W. Griffin and J. S. Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-32(2):236–243, April 1984.

[GW77] S. E. Gerber and B. H. Wulfeck. The limiting effect of discard interval on time-compressed speech. *Language and Speech*, 20(2):108–115, 1977.

[Har90] E. Hardam. High quality time-scale modification of speech signals using fast synchronized-overlap-add algorithms. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 409–412. IEEE, 1990.

[Hej90] D. J. Hejna Jr. Real-time time-scale modification of speech via the synchronized overlap-add algorithm. M.I.T. Masters Thesis, Department of Electrical Engineering and Computer Science, February 1990.

[HLLB86] G. W. Heiman, R. J. Leo, G. Leighbody, and K. Bowler. Word intelligibility decrements and the comprehension of time-compressed speech. *Perception and Psychophysics*, 40(6):407–411, 1986.

[Lee72] F. F. Lee. Time compression and expansion of speech by the sampling method. *Journal of the Audio Engineering Society*, 20(9):738–742, November 1972.

[Lim83] J. S. Lim. *Speech Enhancement*. Prentice-Hall, Inc., 1983.

[LL77] N. J. Lass and H. A. Leeper. Listening rate preference: Comparison of two time alteration techniques. *Perceptual and Motor Skills*, 44:1163–1168, 1977.

[LRRW81] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon. An improved endpoint detector for isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-29(4):777–785, August 1981.

[Mal79] D. Malah. Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-27(2):121–133, April 1979.

[Max80] N. Maxemchuk. An experimental speech storage and editing facility. *Bell System Technical Journal*, 59(8):1383–1395, October 1980.

[ME86] J. Makhoul and A. El-Jaroudi. Time-scale modification in medium to low rate coding. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 1705–1708. IEEE, 1986.

[Min74] F. D. Minifie. Durational aspects of connected speech samples. In S. Duker, editor, *Time-Compressed Speech*, pages 709–715. Scarecrow, 1974.

[ML50] G. A. Miller and J. C. R. Licklider. The intelligibility of interrupted speech. *Journal of the Acoustic Society of America*, 22(2):167–173, 1950.

[MQ86] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34:744–754, August 1986.

[MS62] H. Miedema and M. G. Schachtman. TASI quality—effect of speech detectors and interpolators. *The Bell System Technical Journal*, pages 1455–1473, 1962.

[Neu78] E. P. Neuburg. Simple pitch-dependent algorithm for high quality speech rate changing. *Journal of the Acoustic Society of America*, 63(2):624–625, 1978.

[OFW65] D. B. Orr, H. L. Friedman, and J. C. Williams. Trainability of listening comprehension of speeded discourse. *Journal of Educational Psychology*, 56:148–156, 1965.

[Orr71] D. B. Orr. A perspective on the perception of time compressed speech. In P. M. Kjeldergaard, D. L. Horton, and J. J. Jenkins, editors, *Perception of Language*, pages 108–119. Charles E. Merrill Publishing Company, 1971.

[O'S87] D. O'Shaughnessy. *Speech Communication: Human and Machine*. Addison-Wesley, 1987.

[Por78] M. R. Portnoff. *Time-Scale Modification of Speech Based on Short-Time Fourier Analysis*. PhD thesis, MIT, April 1978.

[Por81] M. R. Portnoff. Time-scale modification of speech based on short-time fourier analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-29(3):374–390, June 1981.

[QM86] T. F. Quatieri and R. J. McAulay. Speech transformations based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34:1449–1464, December 1986.

[Que74] S. U. H. Quereshi. Speech compression by computer. In S. Duker, editor, *Time-Compressed Speech*, pages 618–623. Scarecrow, 1974.

[Rei80] S. S. Reich. Significance of pauses for speech perception. *Journal of Psycholinguistic Research*, 9(4):379–389, 1980.

[Rip75] R. F. Rippey. Speech compressors for lecture review. *Educational Technology*, November 1975.

[RS75] L. R. Rabiner and M. R. Sambur. An algorithm for determining the endpoints of isolated utterances. *The Bell System Technical Journal*, 54(2):297–315, February 1975.

[RSLM88] A. Richaume, F. Steenkeste, P. Lecocq, and Y. Moschetto. Intelligibility and comprehension of French normal, accelerated, and compressed speech. In *IEEE Engineering in Medicine and Biology Society 10th Annual International Conference*, pages 1531–1532, 1988.

[RW85] S. Roucos and A. M. Wilgus. High quality time-scale modification for speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 493–496. IEEE, 1985.

[Sco67] R. J. Scott. Time adjustment in speech synthesis. *Journal of the Acoustic Society of America*, 41(1):60–65, 1967.

[SG72] R. J. Scott and S. E. Gerber. Pitch-synchronous time-compression of speech. In *Conference on Speech Communication and Processing*, pages 63–65. IEEE, 1972. Reprinted in J. S. Lim *Speech Enhancement*, Prentice-Hall, Inc., 1983.

[Sti69] T. G. Sticht. Comprehension of repeated time-compressed recordings. *The Journal of Experimental Education*, 37(4), Summer 1969.

[Too74] H. D. Toong. *A Study of Time-Compressed Speech*. PhD thesis, MIT, June 1974.

[VM65] J. B. Voor and J. M. Miller. The effect of practice upon the comprehension of time-compressed speech. *Speech Monographs*, 32:452–455, 1965.

[WLS84] A. Wingfield, L. Lombardi, and S. Sokol. Prosodic features and the intelligibility of accelerated speech: Syntactic versus periodic segmentation. *Journal of Speech and Hearing Research*, 27:128–134, March 1984.

[WN80] A. Wingfield and K. A. Nolan. Spontaneous segmentation in normal and time-compressed speech. *Perception and Psychophysics*, 28(2):97–102, 1980.

[WRW89] J. L. Wayman, R. E. Reinke, and D. L. Wilson. High quality speech expansion, compression, and noise filtering using the SOLA method of time scale modification. In *23d Asilomar Conference on Signals, Systems, and Computers*, pages 714–717, October 1989. Vol. 2.

[WSB92] L. Wilcox, I. Smith, and M. Bush. Wordspotting for voice editing and audio indexing. In *CHI '92*. ACM SIGCHI, 1992.

[WW88] J. L. Wayman and D. L. Wilson. Some improvements on the synchronized-overlap-add method of time-scale modification for use in real-time speech compression and noise filtering. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(1):139–140, January 1988.

[ZDS68] W. R. Zemlin, R. G. Daniloff, and T. H. Shriner. The difficulty of listening to time-compressed speech. *Journal of Speech and Hearing Research*, 11:875–881, 1968.