Speaker Rate Detection for Voice Mail Applications

by

Scott A. Lehman

Submitted to the Department of Electrical Engineering
and Computer Science in partial fulfillment of the
requirements for the degree of

Bachelor of Science in Electrical Science and
Engineering and Master of Engineering in Electrical
Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1997

Author ....
Department of Electrical Engineering and Computer Science
June 2, 1997

Certified by ....
Christopher M. Schmandt
Principal Research Scientist, Media Lab
Thesis Supervisor

Accepted by ....
F. R. Morgenthaler
Chairman, Department Committee on Graduate Theses

Speaker Rate Detection for Voice Mail Applications

by

Scott Lehman

Submitted to the Department of Electrical Engineering and
Computer Science

June 2, 1997

In Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Electrical Science and Engineering and Master of Engineering in Electrical Engineering and Computer Science

Abstract

Voice mail users often feel frustrated when having to wait through long messages from slow speakers. Increasing the playback speed of the messages helps, but care must be taken so that messages from the fastest speakers remain intelligible. The system can be made even more efficient with the use of a speech rate detector. By analyzing each message with the speech rate detector, the playback speed could then be adjusted for each message, ensuring that all messages can be played back as quickly as possible while retaining intelligibility. While previous research has produced accurate speaker rate detection systems in the laboratory, some problems are encountered when adapting them to the voice mail system.

Thesis Supervisor: Christopher M. Schmandt
Title: Principal Research Scientist, Media Lab

Table of Contents

List of Figures

Chapter 1

Introduction

Although much of the focus and attention on speech processing these days is in the field of speech recognition, there are other important application areas. One such application is speaker rate detection, that is determining how quickly someone is speaking or the rate of information flow. Speaker rate detection shares many of the challenges of speech recognition, notably robustness to different speakers and the variability of the environmental conditions the speaker is in. Similar to speech recognition systems, the control of the environment and speaker dependence decreases, the problem can become substantially harder [1] [6]. But speaker rate detection is free of a very difficult part of continuous speech recognition - determining precisely which phonemes and words are being spoken. A rate detection system then is very similar to the "front end" of a speech recognition system which must reduce the complex speech input into discrete units (such as phonemes or syllables), which in the case of rate detection would then be counted over some time interval rather than looking them up and matching them with entries in the system's vocabulary. By avoiding all the "back end" issues of actually interpreting what is being spoken, it is hoped that a simple and efficient system can be designed for accurate speech rate detection, while avoiding all the cost and complexity of a complete continuous speech recognition system.

The primary interest in speaker rate detection for this thesis lies in voice mail applications, more specifically for retrieval of messages. As a voice mail user, it can be very frustrating to wait for slow speakers to complete their messages, particularly when busy or traveling. This can be remedied to some extent by increasing the playback speed of the messages or by using a more complicated time compression technique that preserves pitch. The problem here is that different people speak at different rates, and we would like to optimize the playback speed of each message individually to match its speaker. If a single playback speed were to be applied to all messages, those messages from people with a very high speech rate may become unintelligible while the recipient may still end up having to wait impatiently through a message from the slowest speaker. By analyzing each voice mail message with an accurate speech rate detection algorithm prior to playback, the voice mail system could then compute the appropriate playback speed that should be applied to each message. This will make the process of getting one's voice mail as efficient as possible.

Speaker rate detection is useful in other applications involving human-computer interfaces as well. Even if the words being spoken and their meaning are not known, speech rate can still provide valuable cues. For example, when designing a system where a computer is dictating a series of instructions or directions to a human, the speech rate of the user can be used as an indicator of his or her understanding of the instructions or ability to process them. A sudden burst of quickly spoken words may mean that the user didn't have time to write something down or just couldn't hear the machine clearly the first time so the instruction could be repeated, while a very slow rate could be interpreted as acceptance by the user.

In this thesis, the syllable is chosen as the basic unit of recognition. This decision is based in part on previous work with automatic segmentation of speech into syllables. That research produced an algorithm with a reported error of only 6.9% syllables missed (relative to the total number of syllables), and 2.6% extraneous segments that did not correspond to actual syllables [3]. The algorithm presented below is a variation on that work and is designed to determine the speech rate, expressed in syllables per second. It has also been adapted to work with the available voice mail system.

For this voice mail application, the speaker rate detection scheme must be tolerant of not only each speaker's characteristics, but also the conditions related to each particular call. Calls from different telephones (or possibly even the same telephone at different times) will be affected differently in terms of speech dynamics, background noise, and distortion. The limited bandwidth of the analog cables will have a filtering effect on the audio that depends on the length of the cable run. Meanwhile, newer digital transmission techniques, such as speech coding for wireless telephone networks, may impart very different qualities on the speech. As time progresses and new technologies come into being, even more variability may be thrown into the mix. Gaining independence over all of these variables may prove to be a daunting task, but it is essential for reliable performance.

Chapter 2

Methods

2.1 Introduction

In order to segment continuous speech into syllables, information must be extracted from the speech waveform. Syllables are generally defined as a period of voiced speech, i.e. a vowel, preceded and/or followed by a consonant. The energy in the segment of voiced speech is referred to as the sonorant energy. Segmentation techniques can try to locate the occurrences of this energy in a variety of ways. One approach is to examine the amplitude envelope or intensity of the time-domain waveform, guided by the thinking that the voiced speech will be characterized by sustained periods of large amplitude while consonants are not as prominent and more transitory in nature. The syllables would then be determined by tracking the peaks and transitions in the envelopes.

An alternate technique for locating syllables would be to break the speech into voiced and unvoiced segments. A number of different indicators can be used to determine the voicing state. Examining the signal energy in a low frequency band from 60 to 400 Hz is one possibility [2]. Voiced segments are also characterized by their self-correlation in time as well as harmonic and formant structure in the frequency domain. Wide-band energy or sharp transients would hint at unvoiced speech. With a reliable voicing detector, the transitions between voiced and unvoiced states, along with periods of silence, would then define the syllable boundaries. Additional syllable boundaries may then be found by momentary drops in amplitude or power during the longer segments of voiced speech [5].

A third approach, and the approach used in this thesis, is to create a "loudness" function, as suggested by Mermelstein [3]. In this case, the short-time power over a selected frequency band is summed to produce the loudness function. The precise frequency band is chosen to fall in areas where the sonorant energy is strongest so that the syllables manifest themselves as peaks in the loudness function. This loudness function is then segmented to locate the syllables.

2.2 Computing the Loudness Function

The loudness function is created using very common signal analysis techniques. The speech sequence is windowed, that window of data is processed, and then the window advances. For each window of speech data, the short-time power spectrum is then computed with a Fast Fourier Transform (FFT). The signal energy could also be obtained with computation on the actual time sequence or filtering, but using the FFT makes it very easy to quickly view the energy in a particular frequency band and to weight certain frequencies more than others. The time domain processing would account for the energy over the entire spectrum which is not necessarily desirable. In this thesis, all data was weighted with a Hamming window. The typical length of the time window used in this work was 16 milliseconds, with a frame increment of 8 milliseconds so that each frame overlaps with half of the previous frame.

Each point of the loudness function corresponds to one windowed frame of the speech data. The value of the loudness function is created simply by summing up the power spectrum over the selected frequency range and applying a logarithmic compression. When analyzing a series of speech recordings, the peak loudness level in each recording was normalized. This sequence is then smoothed with a fifth order, linear phase low pass filter, designed with a cutoff frequency at 200 Hz. The newly smoothed loudness function bears a strong resemblance to the amplitude envelope of the speech segment. Figure 2.1 displays the waveform for a sample phrase and its spectrogram is shown in Figure 2.2. The initial loudness function and the smoothed version are presented in Figure 2.3 and Figure 2.4 respectively.

algorithm to count five syllables when there are only four. By reducing the upper limit of the frequency band, that extra peak in the loudness can be reduced so that the algorithm correctly locates only four syllables, as shown in Figure 3.4.

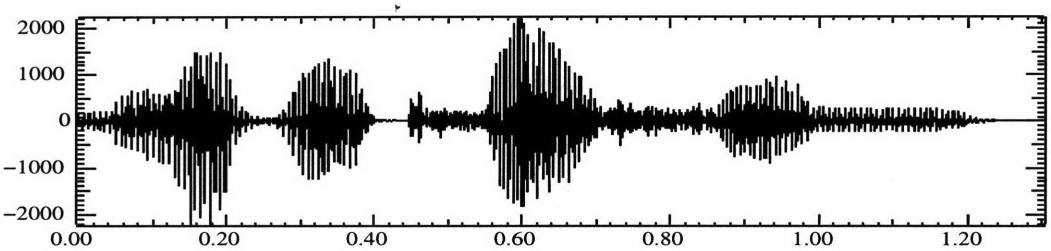

Figure 3.1: The phrase "My vacation."

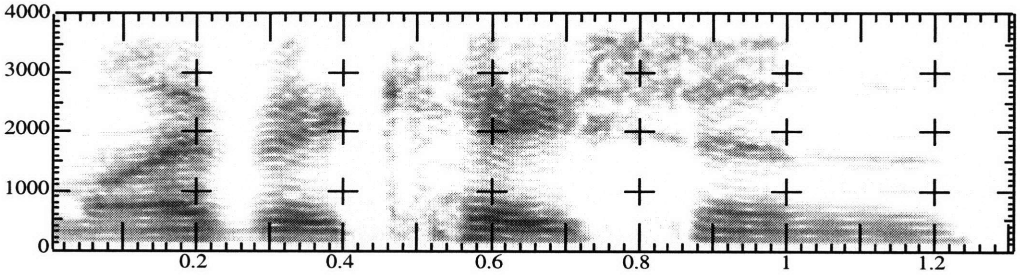

Figure 3.2: Spectrogram of "my vacation."

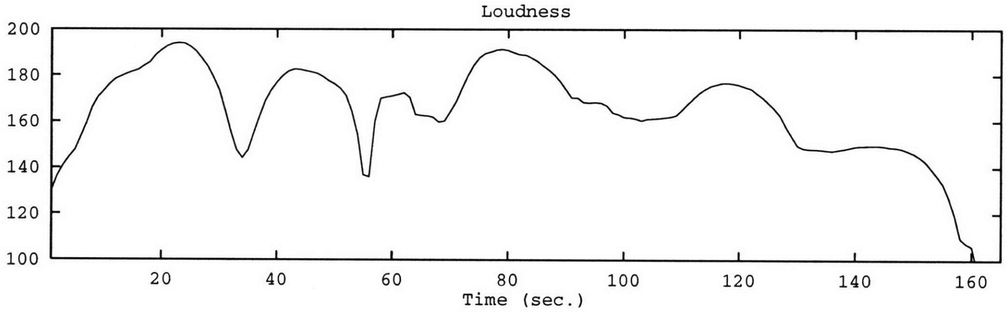

Figure 3.3: Loudness function for "my vacation" with a wide-band analysis.

The final component is to label portions of the loudness function as silence. A silence threshold is defined, and any points in the loudness function that fall below that threshold are then set to the threshold value while larger values are left unchanged. This step is an attempt to discard the variations in the noise of the system or other very quiet sounds which might incorrectly be labelled as syllables in the segmentation process.

2.3 Segmentation with a Convex-Hull

Once the loudness function has been created, the convex-hull algorithm is used for segmenting it into syllabic components. The algorithm proceeds by recursively dividing the loudness function into two smaller components at locations where the differences between local maxima and minima are large.

The convex-hull is defined as the minimum magnitude curve, with respect to the function being segmented, that is strictly monotonically non-decreasing moving from the endpoints of the function to the point of maximum value, behaving much like an ideal peak detector circuit. The largest difference between the convex-hull and the function being segmented is noted. If this difference exceeds a defined segmentation threshold, the function is then segmented into two smaller functions at that point of the maximum difference. Both of the new segments of the function are then segmented by another convex-hull. This process is repeated recursively until the maximum difference between the convex-hull and the function being segmented falls below the defined threshold and is deemed insignificant, or when the only minima are at the endpoints. For the rate detection system, the segmentation threshold is selected so that it reduces the loudness curve into syllabic units. The number of segments produced when the recursion can go no further is the number of syllables. An example of the convex-hull in action is illustrated in Figure 2.5.

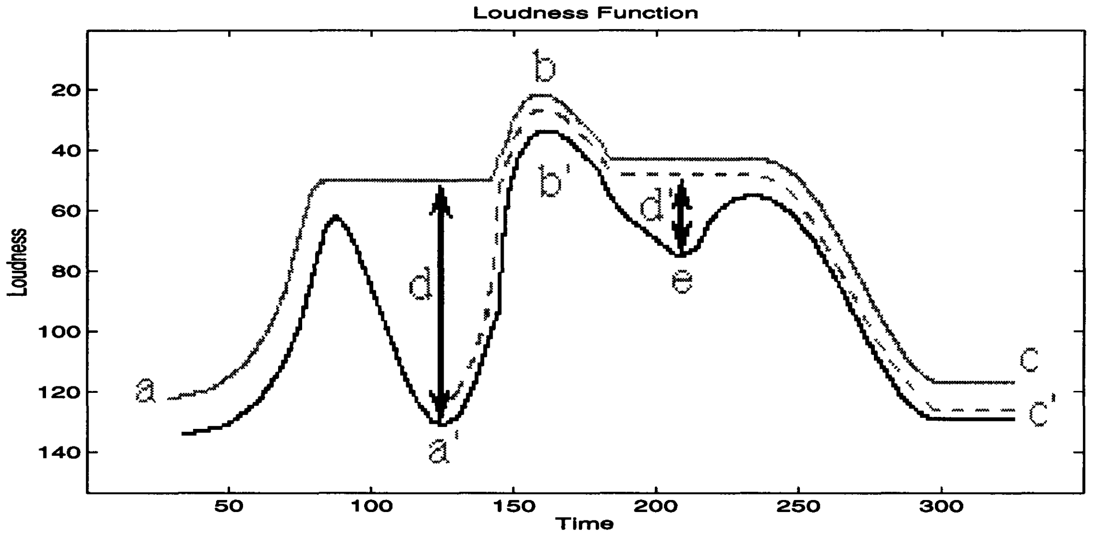

Figure 2.5: An example of the convex-hull applied to an arbitrary function.

The function being segmented is drawn in black in Figure 2.5 and segmentation begins with the segment (a-c). The initial convex-hull is the solid gray line, with the peak value falling at point b. The largest difference between the convex-hull and the function occurs at point a' and has a value of d. This difference is larger than the segmentation threshold so the segment is divided into two - the segment (a-a') and (a'-c'). The first segment does not have a minimum between the endpoints and is not segmented any further. The second does have a local minimum, and its convex-hull is drawn as the dashed gray line. The maximum difference for this segment, d', is located at point e. This difference is below the threshold so the segment is not divided any further.

2.4 Computing the Speech Rate

The simplest method of computing the speech rate is to take the total syllable count and divide by the message length in seconds, however this is not the ideal method. A global rate such as that will be strongly affected by long pauses or lack thereof. In general, the speech rate over the course of a message will not be constant. Long pauses that reduce the global speech rate would suggest an increase in the playback speed which may make certain portions of the message unintelligible. It is most important to find the portion of the message with the highest rate over some shorter time interval and determine the playback speed on that rate.

Computing the maximum rate over too short of a region can be problematic as well. A very short window would be likely to fall on a large, polysyllabic word. Such words tend to have a lower information density than monosyllabic words. For example, people very often drop a syllable when saying "probably" or "everybody" and yet the words can be fully understood. Polysyllabic words can be sped up towards unintelligibility more so than monosyllabic words and still be interpreted properly. Also, since any interval was only permitted to have a integer number of syllables, results for shorter intervals will have much more quantization of the data points. The maximum speech rate reported in this thesis is based on a five second interval.

Chapter 3

Dealing With Sources of Errors

3.1 Introduction

Ideally, all the peaks in the loudness function will be strictly due to the sonorant energy in the speech. However there are other audio elements present in voice mail messages that have enough energy to create large peaks that may mistakenly be counted as syllables. Some of these errors in the loudness function are quite common, but can be controlled to some degree by carefully refining the process of computing the loudness function. But as it turns out, not all the errors are quite that simple to deal with and will require some additional work to compensate for.

3.2 Consonants

The English language is made up of several groups of phonemes. The vowels and their sonorant energy are only one of these groups, and any algorithm must be prepared to handle energy due to other phoneme families. In the family of consonants, the fricatives, affricates, and plosives are the most significant contributors to false peaks in loudness.

When blindly computing the loudness function across the entire spectrum, strong fricatives can produce peaks in the loudness comparable to the peaks from the sonorant energy. Notable examples are the hard consonants like the /k/ in call (plosive) or the /tʃ/ in church (affricate). The energy in these and some other fricatives is concentrated in the higher frequencies, so relying less on this frequency band when computing the loudness function can help prevent errors. Figure 3.1 through Figure 3.4 present an example of this problem with the phrase "my vacation" and how lowering the upper bound of the frequency band improves things. The loudness function shown in Figure 3.3, which uses a wide frequency band, has a significant peak due to the /k/ in vacation, which caused the algorithm to count five syllables when there are only four. By reducing the upper limit of the frequency band, that extra peak in the loudness can be reduced so that the algorithm correctly locates only four syllables, as shown in Figure 3.4.

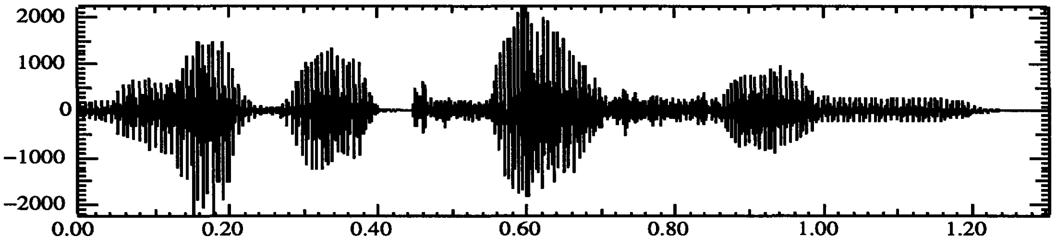

Figure 3.1: The phrase "My vacation."

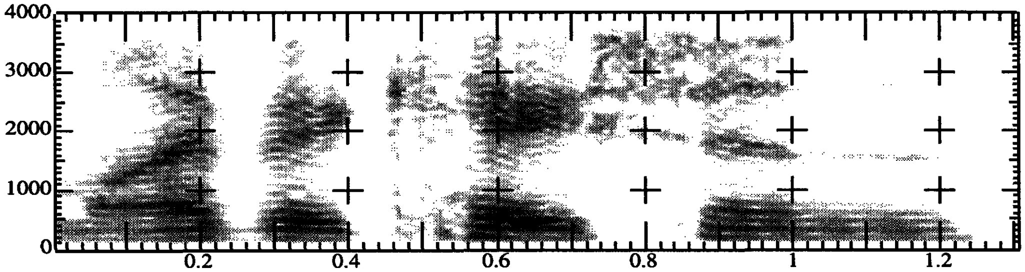

Figure 3.2: Spectrogram of "my vacation."

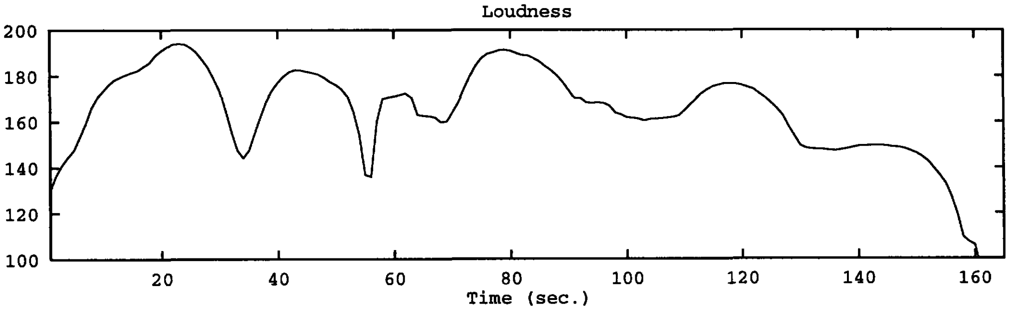

Figure 3.3: Loudness function for "my vacation" with a wide-band analysis.

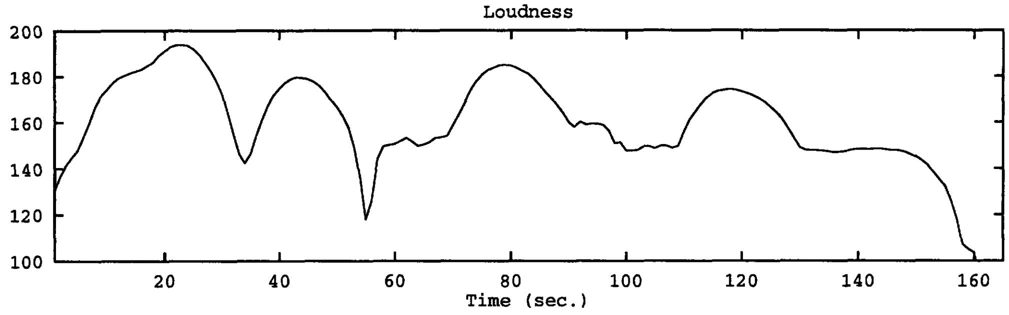

Figure 3.4: Loudness for the "my vacation" using a narrower-band analysis.

Unfortunately, not all consonants are created equal. While peaks from the hard /k/ can generally be identified and eliminated, other plosives such as the /p/ in put and the /t/ in too are quite a bit harder to deal with. The energy of plosives such as these is more evenly distributed across the entire frequency spectrum and are not simply isolated in a small frequency band. Figure 3.5 displays the time waveform for the word "chat" where the closing plosive is heavily emphasized. The spectrogram in Figure 3.6 shows the plosive's wide-band energy distribution while the opening affricate's energy is concentrated in the high frequencies. With a wide-band analysis, the loudness function for the word contains an extra peak for both the affricate and plosive, producing two false syllables as shown in Figure 3.7. The loudness function that results by narrowing the frequency band in the analysis is presented in Figure 3.8. In a manner similar to the previous example, the loudness peak due to the affricate is reduced so that it is not mistaken for a syllable, but the effect on the plosive is minimal. The algorithm still produces the incorrect syllable count. Plosives like this do seem to happen less often than the fricatives and affricates however.

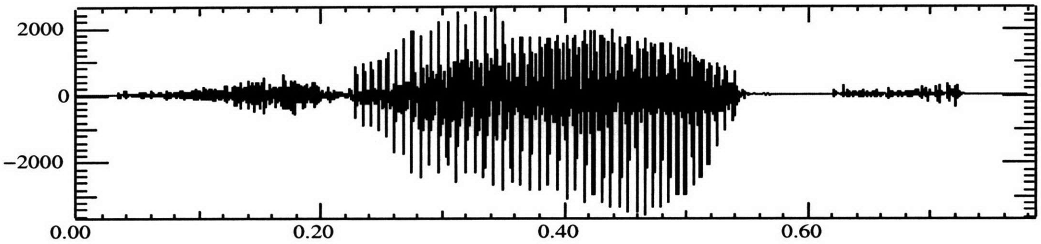

Figure 3.5: The mono-syllabic word "chat."

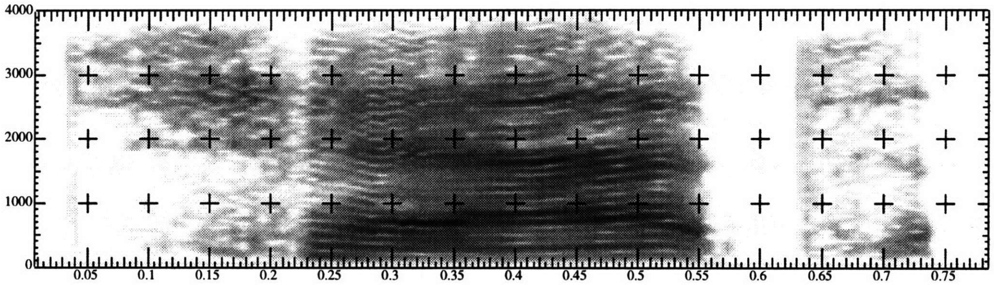

Figure 3.6: Spectrogram of the word "chat."

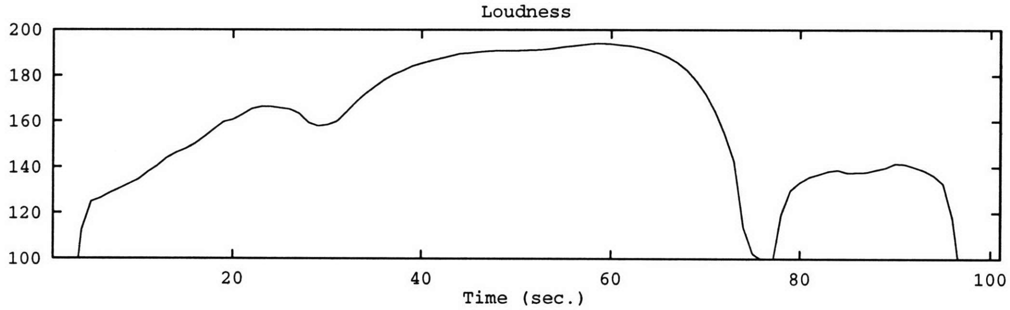

Figure 3.7: Loudness function for "chat" with a wide frequency analysis band.

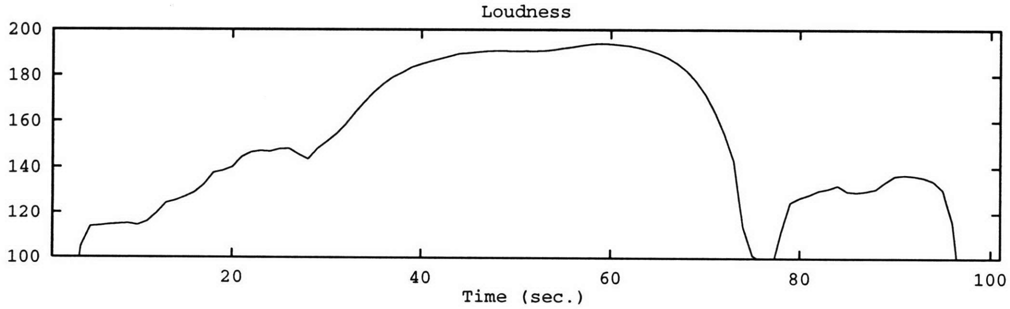

Figure 3.8: Loudness function of "chat" generated using a smaller frequency band.

3.3 Breaths

One of the frequent non-speech sounds that appear in voice mail messages is breathing. Breaths are similar to the plosives discussed in the previous section in that their energy is typically very wide-band and not easily isolated by ignoring a specific frequency band. A brief speech segment that includes a deep inhalation is shown in Figure 3.9. The spectrogram in Figure 3.10 displays the wide-band nature of the breath, although the energy is a bit more concentrated in the higher frequencies. For this particular case, the loudness of the breath, shown in Figure 3.11, has enough variation to be counted as multiple syllables.

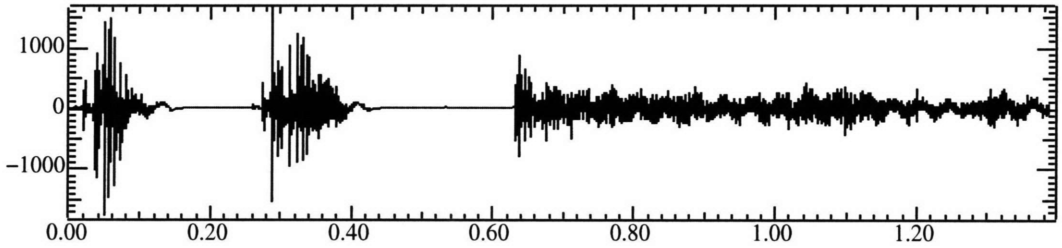

Figure 3.9: "Birthday" followed by an inhalation.

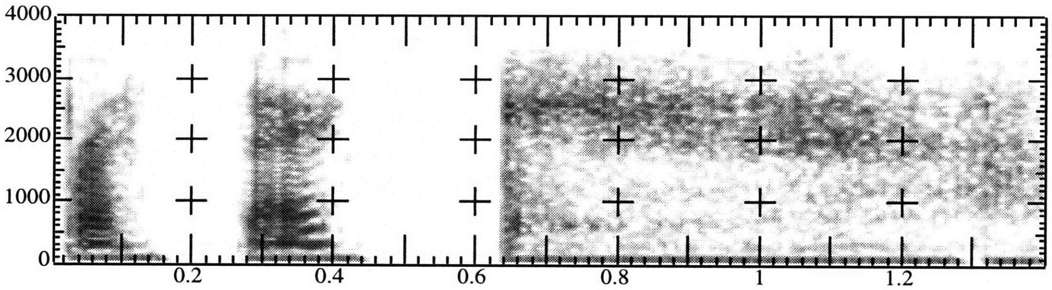

Figure 3.10: Spectrogram of "birthday" and the inhalation.

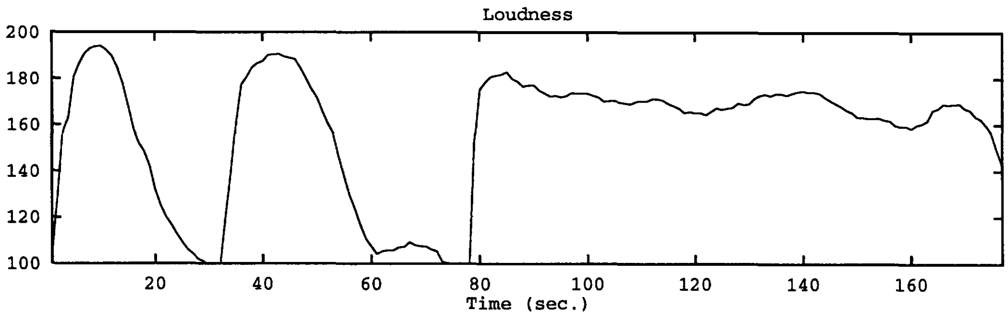

Figure 3.11: Loudness function of "birthday" and the inhalation.

3.4 Callers Hanging Up

Working with the telephone system brings in new sources of noise that wouldn't be a problem in other applications. Perhaps the most significant such noise comes from the callers hanging up the phone after completing their message. As the handset is returned to the cradle of the phone, a series of "clicks" and "clunks" may be recorded while the line is still open. A segment of a message ending with the words "take care, bye," followed by the hang up noise is shown in Figure 3.12, where the first three bursts are the speech. Its spectrogram is presented in Figure 3.13

As can be seen in the loudness function for this example (Figure 3.14), these clicks tend to happen in rapid succession, which may cause the algorithm to report an excessively large speech rate. However, out of the fifteen messages containing hang ups that were used to generate the results in the following chapter, the noise affected the algorithm's results in only one case. (The ten other messages did not have any noise due to hanging up. It is unknown at this time if those files were edited by a previous researcher.) This suggests that this problem is not very severe, which is reasonable considering that people do tend to speak more slowly as the complete their messages and there is usually a period of silence between the last words and hanging up which will reduce rate of energy bursts. The algorithm was later modified to reduce the chances of the error actually happening by simply ignoring the last few seconds of each message. This solution is not ideal since the length of time over which the noise is generated will vary from one call to the next. Perhaps a special filter or modification to the algorithm could provide a better solution.

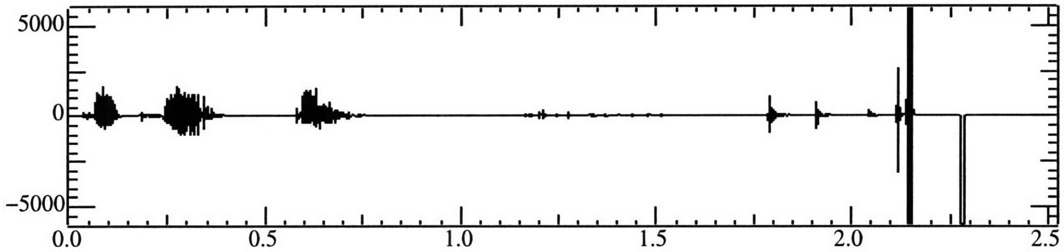

Figure 3.12: The words "take care, bye" followed by noise from hanging up

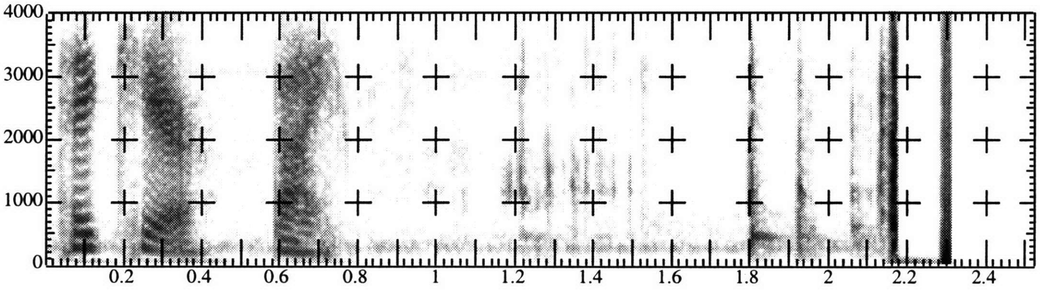

Figure 3.13: Spectrogram of "take care, bye" and hang up noise.

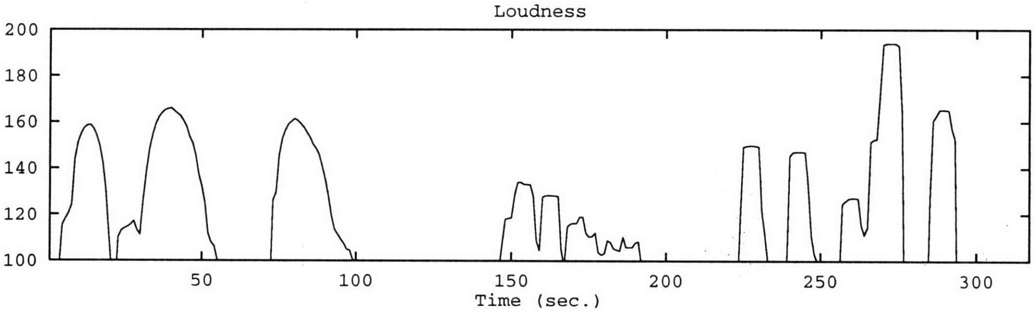

Figure 3.14: Loudness function for "take care, bye" and hang up noise.

3.5 Nullifying Loudness Peaks

Rather than trying to select the frequency analysis band to prevent loudness peaks due to consonants and plosives, other means could be used to selectively nullify the loudness peaks due to those sounds during the segmentation process. One possibility is using a zero crossing detector. The zero crossing density is a crude indicator of high frequencies which might reveal the location of consonants. This way, if the desired frequency band used for capturing the sonorant energy does overlap with the energy of the consonants, those consonants could be found without restricting the sonorant energy range.

Some work was done with a zero crossing detector to nullify peaks due to breaths and certain plosives, but it was not used as it was quite difficult to keep the detector adaptive enough to cope with different speakers. In one particular case, the voice mail message being analyzed had significant distortion and the zero crossing densities were quite abnormal. It was also very important to eliminate any DC component in the speech segment or else the zero crossing densities computed would be lower than the actual values. The simplicity in the implementation of the zero crossing detector makes it an attractive option, but a significant amount of complexity would be needed to make it a reliable tool. Other methods for finding the consonants and plosives that were not explored in detail would be looking for even distributions of energy across the entire spectrum or brief bursts of energy with a uniform onset time that accompany plosives.

The segmentation algorithm could also be enhanced by enforcing a minimum syllable length. It was found that when working with a wide range of speech rates that many of the peaks due to syllables were themselves quite brief, and it becomes difficult to accurately discriminate between actual and false syllables on the basis of segment length alone. The idea may be worth investigating further however.

3.6 Other Considerations

Selecting the frequency range should also take the effects of the telephone system into account. The filtering effects of the system can reduce the frequencies present in the speech. This is a concern since each message may be affected differently. Given the mix of digital and analog hardware these days, the true bandwidth of the telephone channel can vary which could cause problems if the rate detection algorithm relies heavily on frequencies may be present in some messages, but not all of them. A typical medium-range telephone channel (180-725 miles) has a bandwidth near 2400 Hz, somewhere in the range of 300 to 2700 Hz [4]. It would be desirable for the algorithm operate well in the worst case, but even if it does, the coloration induced by the phone system could still be an issue since the magnitude of the transfer function still varies to some degree. The only way to truly eliminate this variable would be to have the phone system perform some type of channel negotiation at the beginning of a call to determine the channel characteristics, much like modems do.

For the particular application of adjusting the playback speed of voice mail messages, if significant errors in the rate calculation are unavoidable it would be preferable to have the rates in error be larger than the actual values. If the rates were reported as being lower, it would imply using a higher playback speed which may make the message unintelligible. Calculated rates that are too large will merely cause the message to be played slower than necessary, and while not as efficient as possible, it will keep the speech understandable.

Chapter 4

Experimental Results

4.1 Introduction

A total of 25 different voice mail messages were used to evaluate the performance of the rate detection program. The particular voice mail system used stores all messages on disk in mu-law encoded format with a sampling rate of 8000 Hz. The messages were also subject to power normalization prior to the rate analysis. No more than three files were from any one speaker, and only three of the speakers were represented with more than one message. Twelve of the messages were spoken by males, and thirteen by females. The messages were chosen from a set to have a variety of different recording environments, noise levels, and distortion levels, as well as higher quality, clean recordings. All messages were manually segmented to locate all the spoken syllables and compute the actual speech rates. The exact locations of the syllables were strictly enforced when determining the algorithm's syllable counting accuracy. When computing the loudness function, it was found that using a frequency band starting in the range of 550 to 650 Hz. with an upper bound between 2500 Hz. and 3000 Hz. seemed to work well. The exact range used to produce the results below was from 550 Hz. to 2700 Hz. These values were chosen based on an extensive series of tests with the algorithm. The length of time required to for the algorithm to perform speech rate analysis was roughly 140% of the actual message length in seconds on a Sparc 10 workstation.

4.2 Results

The ideal case is to have all the data points fall on a line through the origin with the slope of one. The best results obtained, in terms of the total mean squared error and the mean squared error of the points that fall below the ideal line (based on the thought that the low data points which lead to faster playback rates are more critical), is shown in Figure 4.1.

The results do exhibit the desired correlation to some extent, but there is a substantial variance in the computed rate for messages with the same or very similar speech rates, and the output range is considerably less than the input range. Although the program may be of some use as is, the distribution is really not tight enough for critical applications.

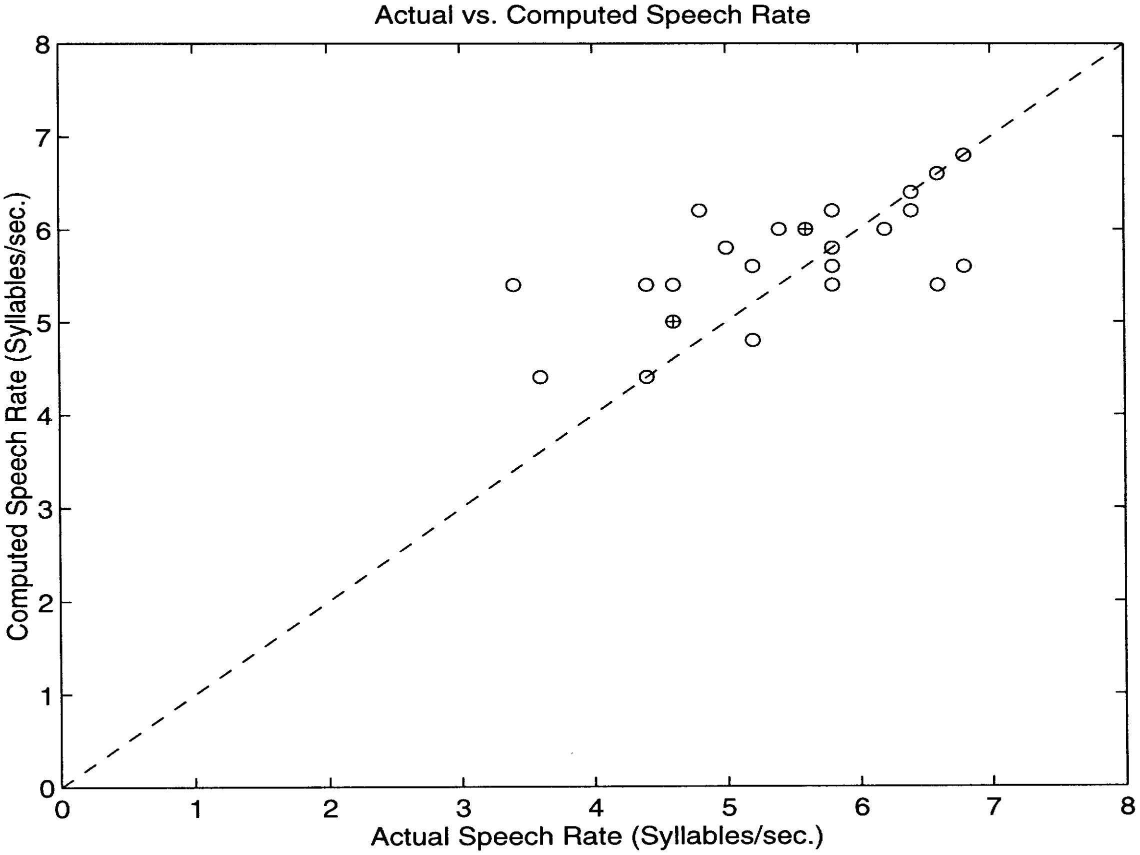

Figure 4.1: The experimental results

In order to achieve something this close to the desired results, the silence threshold was essentially removed entirely. This means that the loudness function may have contained many more artifacts due to noise. For some of the voice mail messages, this had very little or no effect as far as counting too many syllables since there was a smooth and level noise floor, but in other cases it did cause the recognition of false syllables. It turns out to be rather difficult to apply a silence threshold that works reasonably well for all messages. In order to prevent syllables from falling below the silence threshold, a number of peaks in loudness not due to syllables had to be accepted. The lack of consistency in the variability of the noise floor between messages is most likely a substantial reason for the large variance in the computed speech rates.

The variance in the output of the rate detector can also be seen by looking at the algorithm's performance as a syllable counter. The performance is rather poor. Missed syllables are generally less than 20% of the actual, but extra syllables ranged from zero to 50%, and in a couple of cases, beyond even that. So in terms of total syllable tallies for each message, the results in almost every case exceeded the actual count, often substantially. This is not to say that the program is strictly terrible for syllable counting purposes however. With more conservative parameters, the missed and false syllable counts can be tamed and the total syllable counts can be kept between about 80% and 110% of the actual count. But with those settings, the computed speech rates were notably worse than shown above. Computing the maximum short-time speech rate is not necessarily the same as counting all the syllables. An extremely accurate syllable segmentation algorithm should perform well when adapted for rate detection, but a good rate detector only needs to work well in just one segment of the message. It appears to be very difficult to achieve the best syllable counting performance simultaneously with the best rate detection performance with the current implementation of the algorithm.

Even with better management of the silence threshold, some of the voice mail messages are just very difficult to examine accurately. The most significant outlier in Figure

4.1 is particularly tricky because of the noise present. The noise isn't from the phone system, but rather the speaker. Besides a hard consonant and plosive, a number of other sounds and clicks resulting from lip smacks or other actions are clearly spotted. A rapid succession of such sounds is enough to throw the rate estimate way off the mark.

The next most significant outlier has a similar problem as well. Although two clicks could be removed with a properly place silence threshold, a breath causes two false syllables, and fricatives are more prominent in loudness than many of the actual syllables. This is a clear case where simply restricting the high frequencies in the analysis band is not enough.

Moving on to the pair of most significant outliers that fall under the desired values, a different problem is found. The speech is quite fast and the cues to locate the syllables are not as obvious. This appears to be at least partly due to the fact that the speaker is talking so fast that the existing consonants aren't getting enough emphasis to introduce valleys in the loudness (and the hard fricatives that produce peaks in the loudness are almost nonexistent). The groupings of adjacent syllables with vowel sounds become difficult to differentiate. Perhaps the most obvious question is, what happens when you increase the time resolution with a shorter window? This can indeed make some gentle dips in the loudness more significant and introduce some new ones. But this effect is seen pretty much throughout the loudness curve whether they actually correspond to syllable boundaries or not. The configuration parameters for the algorithm become much more sensitive, sometimes resulting in more variance in the computed rates.

This algorithm is not sophisticated enough to find these subtle syllable boundaries right now. The solution may be to take a closer look at the spectrum rather than just summing up the power. The vowel changes should be reflected by the spectral trajectories and the formants and looking for where the sonorant energy is concentrated and how that concentration changes over time could provide some meaningful information to improve the syllable recognition. Of course some measures will need to be taken so that the dipthongs are not mistakenly counted as more than one syllable. Perhaps it is possible that the spectral trajectories of the common dipthongs could be programmed into the algorithm to prevent this while improving the overall syllable detection.

Obviously, any algorithm will have limits on just how high the speech rate can be and still be accurately determined. Eventually the time resolution won't be fine enough to pick out all of the necessary cues, if the cues are still present. However, I think it is clear that this algorithm as currently presented is not hitting that wall.

Chapter 5

Discussion

5.1 Comparison to Previous Work

When the algorithm discussed above is finely tuned to certain messages, excellent performance means that the missed and false syllables are under twenty percent of the actual syllable count. Yet Mermelstein has reported only 6.9% syllables missed and 2.6% incorrectly found syllables. That is a massive disparity, for which there are a number of likely causes.

First of all, the actual speech rates that Mermelstein was using with his algorithm is not exactly known. He reported that his algorithm was tested on "continuous, reading-rate speech," but no specific measurements were stated so it's hard to judge if the voice mail messages used in this work are comparable. Almost certainly, at least some of the speech rates found in the voice mail would be greater than what he described. If the voice mail messages with the highest speech rates were not used to test the algorithm in this thesis, it could have been optimized with a different set of parameters that may have improved the performance at least by a little.

Aside from the speech rate that Mermelstein's work dealt with, the manner in which the subjects were speaking could also be a significant factor. If Mermelstein's speech segments were taken from people reading aloud (as implied by his description of the speech rates involved), the intonation and pronunciation of the speech could be very different than what will be found on a voice mail system. When someone is given a document to read, particularly if sitting in front of a microphone and knowing the he or she is being recording for a specific purpose, the chances are good that the speech will not be very natural. With a voice mail system, the speakers are more likely to be in a comfortable environment since they have become accustomed to leaving messages for people. And callers are also talking off the top of their heads without any written cues. That alone may allow people to speak much more quickly since no optical recognition and reading delay are involved. Perhaps this could be alleviated a bit if the voice mail system were to simply inform the callers that their message would be processed by a computer and to please speak clearly. This might make the caller a little more conscious about how he or she is speaking, resulting in messages that are easier to analyze accurately.

Another possibly important difference between Mermelstein's work and this thesis could be the recording environment, which again is not documented. Presumably, the speech recordings took advantage of the luxury of a controlled environment. Using the same hardware system for the recordings will probably keep the noise level in the recordings relatively consistent. The speech dynamics could probably be controlled more and distortion in the speech could be all but eliminated. With the telephone system however, these advantages are lost. The real world is very different than the laboratory.

The actual recording hardware may also be very important. The proximity of the microphone in a telephone handset to the speaker’s mouth may well capture information that wouldn’t be seen if the microphone were place several feet away. For example, with a telephone, the speaker's breathing tends to create turbulence as the air flows across of the mouthpiece, but what if the mouthpiece wasn’t so close? Breaths were very prominent in almost all of the voice mail messages examined in this work and they were a contributor to the computed maximum speech rate in eleven of the messages. It could be argued that a breath alone could be counted as a syllable as it does carry some information about the speech rate and typically does not coexist with spoken words. But even so, in several instances a single inhalation or exhalation had a very irregular loudness and was segmented as more than one syllable. The microphone’s proximity may also make other sounds such as lip smacks, clicks produced by the mouth when opening or closing, and wetting ones lips much more prominent in the loudness than they would be with the microphone further away. Again, the most significant outlier in the data plotted above ended up there because of sounds like these. Incorporating a declicker into the rate detection algorithm or for pre-processing the audio files may eliminate these sounds and perhaps those of the phone hanging up as well.

There is one more difference between this voice mail system and Mermelstein's system, but it is likely inconsequential. The difference is the sampling rate of the audio. Although not explicitly listed, Mermelstein was working with audio sampled greater than 8 kHz since he mentions weighting frequencies above 4 kHz, which is the Nyquist frequency in this voice mail system. The higher sampling rate would surely provide some additional information, but it is not going to close the performance gap between Mermelstein's implementation and the one presented here. However the bandwidth differences in the hardware may have a measurable effect. As more things go digital, the telephone system's speech quality should improve, but it will take years, if ever, before everything can be upgraded.

5.2 Alternative Implementations

In cases where real-time operation is required, the above algorithm would need to be altered. Even when the computation can be completed in a duration shorter than the message length, the convex-hull segmentation algorithm requires that the entire speech segment be stored in memory before running. A new segmentation algorithm could be used or the speech could be stored into a buffer until a long enough pause is detected, and then that segment could be analyzed in quasi-real-time.

One possible thought for future work might be intelligently deciding which portions of messages are more critical for the rate analysis. For example, a number of voice mail messages tended to have high speech rate at the beginning of a message, and then the speech slows down a bit. This may be more practical when there is a very accurate rate detection scheme that is very computationally intensive.

Perhaps rate detection could be combined with a speaker recognition system so the system could then look up the typical speech rate for a frequent caller. Then the rate could be altered over time by the voice mail recipient as well so if one individual generally has nothing important to say, his messages can be played faster.

It might also be interesting add a little twist on the message playback system, such as adjusting the playback speed while the message is be heard. Or it may be relatively simple and effective to automatically edit out the pauses in a message prior to playback to streamline things a bit more.

5.3 Conclusions

Developing a rate detection algorithm for processing voice mail turned out to be a tricky problem that requires more than a relatively simple algorithm. There is quite a bit of variability between speakers and in the hardware that makes up the telephone network as well. And unfortunately there seems to be a lack of publications that deal directly with speech rate detection algorithms. Before proceeding with future work, several items should be seriously addressed.

1. There needs to be a robust mechanism for dealing with the various clicks, breaths, and other sounds so they are not incorrectly counted as syllables.

2. The algorithm needs to be adaptive to the variation in both the noise level of the phone system and the dynamics of the speech. The lack of a silence threshold may not be

a significant factor in some cases, but tailoring it to all messages effectively will reduce the variance of the computed rates.

and 3. The expected maximum speech rate should be defined beforehand. This will determine the required complexity of the algorithm, and even if it is lower than real world values, it may be easier to enhance a system that performs well at these rates to handle higher rates.

References

[1] Jean-Claude Junqua and Jean-Paul Haton. *Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Boston, Massachusetts, 1996.

[2] Wayne Lea, ed. *Trends in Speech Recognition*, Prentice-Hall, Englewood Cliffs, New Jersey, 1980.

[3] Paul Mermelstein. Automatic segmentation of Speech into Syllabic Units. *Journal of the Acoustical Society of America*, 58, pages 880-883, 1975.

[4] John G. Proakis. *Digital Communications*, McGraw-Hill, New York, 1983.

[5] Arthur Stowe. Segmentation of Speech into Syllables. *Journal of the Acoustical Society of America*, pages 806-807, 1963.

[6] William Woods, ed. *Computer Speech Processing*, Prentice-Hall, Englewood Cliffs, New Jersey, 1983.