MACH1: NONUNIFORM TIME-SCALE MODIFICATION OF SPEECH

*Michele Covell, Margaret Withgott,** *and Malcolm Slaney*
Interval Research Corporation, 1801 Page Mill Road, Bldg. C, Palo Alto, CA 94304 USA

ABSTRACT

We propose a new approach to nonuniform time compression, called *Mach1,* designed to mimic the natural timing of fast speech. At identical overall compression rates, listener comprehension for Mach1-compressed speech increased between 5 and 31 percentage points over that for linearly compressed speech, and response times dropped by 15%. For rates between 2.5 and 4.2 times real time, there was *no* significant comprehension loss with increasing Mach1 compression rates. In A–B preference tests, Mach1-compressed speech was chosen 95% of the time. This paper describes the Mach1 technique and our listener-test results. Audio examples can be found on http://www.interval.com/papers/1997-061/.

1 LINEAR TIME COMPRESSION

Time-compression techniques change the playback rate of speech without introducing pitch artifacts. However, when linear-compression techniques are used, human comprehension of time-compressed speech typically degrades at compression rates above two times real time [1]. These degradations are not due to the speech rate *per se:* Comprehension of linearly compressed speech often breaks down above 225 to 270 words per minute (wpm) [2], which is well below the rates at which long passages of natural speech are comprehensible (up to 500 wpm) [3].

Instead, the incomprehensibility of time-compressed speech is due to its unnatural timing. Mach1, described in Section 2, is an alternative to linear time compression. Mach1 compresses the components of an utterance to resemble closely the natural timing of fast speech. Section 3 describes our test of comprehension and preference levels for Mach1-compressed and linearly compressed speech. In Section 4, we draw our conclusions.

2 MACH1 TIME COMPRESSION

Mach1 mimics the compression strategies that people use when they talk fast in natural settings. We used linguistic studies of natural speech [4,5] to derive these goals:

- Compress pauses and silences the most
- Compress stressed vowels the least
- Compress unstressed vowels by an intermediate amount
- Compress consonants based on the stress level of the neighboring vowels
- On average, compress consonants more than vowels

Also, to avoid obliterating very short segments, we need to avoid overcompressing already rapid sections of speech.

Unlike previous techniques [6,7], Mach1 deliberately avoids categorical recognition (such as silence detection). Instead, as illustrated in Figure 1, it estimates continuous-valued measures of local emphasis and relative speaking rate. Together, these two sequences estimate what we call *audio tension:* the degree to which the local speech segments resist changes in rate. High-tension segments are less compressible than low-tension segments. Based on the audio tension, we modify the target compression rate to give local target compression rates. We use these local target rates to drive a standard, time-scale modification technique (e.g., synchronized overlap-add [8]).

2.1 Local-Emphasis Measure

We use the *local-emphasis measure* to distinguish among silence, unstressed syllables, and stressed syllables. Emphasis in speech correlates with relative loudness, pitch variations, and duration [9]. Of these, relative loudness is the most easily estimated.

2.1.1 Estimating local energy

To estimate local emphasis, we first calculate the local energy. We simply use the frame energies from the spectrogram used in speaking-rate estimation (see Section 2.2).

2.1.2 Normalizing by the local energy average

Emphasis is indicated more by relative loudness than by absolute loudness. So, we normalize our local energy by the average energy. We use a single-pole low-pass filter (τ = 1 sec) to estimate the energy average.

2.1.3 Reducing the dynamic range

The variations of the local relative energy are not linearly related to our goal: controlling the segment-duration variations to mimic those seen in natural speech. In the data that we collected, the local relative energy within emphasized vowels averages around 4.4, with variations from 1.6 to as high as 40. In contrast, the relative

* Current address: Electric Planet, 3200 Ash, Palo Alto, CA 94306.

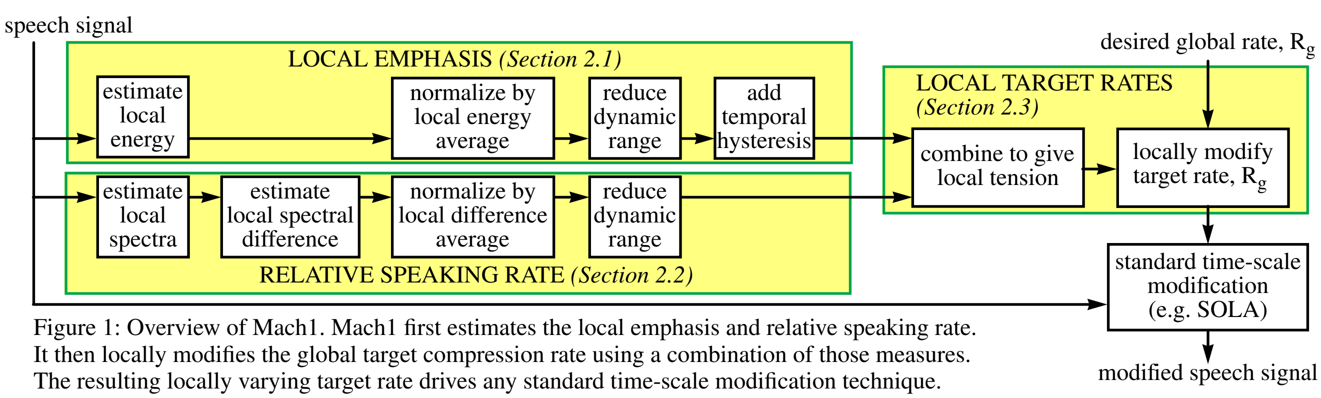

Figure 1: Overview of Mach1. Mach1 first estimates the local emphasis and relative speaking rate. It then locally modifies the global target compression rate using a combination of those measures. The resulting locally varying target rate drives any standard time-scale modification technique.

variability in the compression rates of stressed-vowel durations observed in naturally fast speech is closer to 22% [4]. At the same time, the local-energy variations between unstressed vowels and pauses are less than the compression-rate variations for those segments seen in natural slow and fast speech [4,10].

Therefore, we estimate the *frame emphasis* by reducing the dynamic range of the large-relative-energy segments (the emphasized vowels) and by expanding the dynamic range of the small-relative-energy segments (the unemphasized vowels and the pauses). Currently, our remapping function is hard limiting (to below 2) followed by a square-root function.

2.1.4 Adding temporal hysteresis

Speech perception and production include temporal grouping effects. In American English, all segments in stressed syllables tend to be less variable than those in unstressed syllables [5]. This observation implies that unvoiced consonants in a stressed syllable need to be treated as emphasized, even though their frame-emphasis values are low: The reduction of the consonants is controlled more by the stress of neighboring vowels than by the signal characteristics of the consonant itself.

Similar temporal-grouping effects are present in pauses within speech. Long interphrase pauses can be reduced to 150 msec with little effect on comprehension [6]. Below 100 to 150 msec, further interphrase pause compression causes false pitch-reset percepts.* Thus, our intention is to treat silences near voiced speech as speech and to compress heavily silences outside of this range.

To account for these temporal grouping effects, we apply a tapered, temporal hysteresis to the frame emphasis to give our final local-emphasis estimates. Our hysteresis extends the influence of each frame-emphasis value by 80 msec into the past and 120 msec into the future. To minimize discontinuities in the local emphasis, we taper the hysteresis, using a triangle function to extend each frame-emphasis value into the past and future. We then find the maximum tapered future (or current) frame-emphasis value and the maximum tapered past (or current) frame-emphasis value. The local-emphasis value is the average of these two tapered maxima.

2.2 Relative Speaking-Rate Measure

We estimate the speaking rate to avoid overcompressing, and thereby obliterating, short segments in already rapid speech. True speaking rate is difficult to measure. We can, however, easily compute measures of acoustic-variation rates, which covary with speaking rate. By lowering our compression during transitions, we effectively lower the compression of rapid speech. This approach also has the advantage of preserving phoneme transitions, which are particularly important for human comprehension [11]. Instead of transition labels, we use relative acoustic variability to modulate the compression rate, thereby avoiding categorical errors.

2.2.1 Estimating local spectra

Our estimate of relative acoustic variability starts with a local spectral estimate. We use the spectral values from a preemphasized narrow-band spectrogram, with a frame length of 20 msec and a step size of 10 msec.

To avoid unreliable estimates in low-energy regions, we set each frame whose energy level is below a dynamic threshold to the previous frame's values. Our dynamic threshold varies linearly with the local energy average (described in Section 2.1.2); for the results reported here, we set the threshold such that frames with energy levels below 4% of the local average are reset.

2.2.2 Estimating local spectral difference

Intensity-discrimination studies [12] suggest that human perception of acoustic change is closely approximated by $\log(1 + \Delta I/I)$, where I is intensity. We therefore use the sum of absolute log ratios between the current and the previous frames' values to estimate the local spectral difference.

To avoid overestimating the spectral difference due to simple (scalar) changes in loudness, we normalize each frame's values by that frame's total energy level prior to taking the absolute log ratios. To avoid overestimating the spectral difference due to unreliable values in low-energy frequency bins, we sum over the most energetic bins only. Currently, we sum over the bins within 40 dB of the maximum current-frame value.

2.2.3 Normalizing by the local spectral-difference average

Different speaking styles and different recording environments introduce wide deviations in our absolute spectral-difference measure. To avoid being unduly affected by these variables, we normalize our difference measure by the local average difference.

Guided by informal listener tests, we estimate the local average difference using a weighted average. The average is weighted by the local-emphasis measure (computed in Section 2.1). We compute the average using a single-pole, low-pass filter ($\tau = 1$ sec).

2.2.4 Reducing the dynamic range

The variations of the relative spectral difference overestimate the upward variations in relative speaking rate. The upper 1% of the weighted spectral-difference values range from 4 to 10 times the average. In contrast, segmental-rate variations in natural speech remain well below this range [4]. Therefore, this step estimates the *relative speaking rate* by simply hard-limiting the relative spectral difference to below four times the average.

2.3 Local Target Compression Rates

The local-emphasis and relative speaking-rate measures depend purely on the audio signal that we plan to modify: They can be computed as the signal is being recorded. What remains is to combine these two measures together, to get a single measure of the compressibility of the underlying speech, and to then combine that compressibility measure with the listener's target compression (or expansion) rate.

2.3.1 Computing audio tension

We compute audio tension from local emphasis and relative speaking rate using a simple linear formula:

$$\gamma(t) = \alpha((E(t) - M_E) + \beta(S(t) - M_S)) \quad ,$$

where $\gamma(t)$ is an audio-tension value; $E(t)$ and $S(t)$ are local-emphasis and relative speaking-rate estimates; M_E and M_S are their mean values; $\alpha > 0$ is a constant-valued coefficient; and β is a constant-valued coefficient, with $\beta > 0$ for time compression and $\beta < 0$ for time expansion. For the results reported here, we set $\alpha = 1/2$ and $\beta = 1/4$. For simplicity's sake, we set M_E and M_S to the prior estimates of the means, derived from a large set of speech samples: specifically, $M_E = 0.7$ and $M_S = 1.0$.

Thus, the audio tension increases as the local emphasis increases, from low tension (large compressions or expansions) in regions of silence to high tension (small compressions or expansions) in stressed segments. For time compression, the audio tension increases as the relative speaking rate increases, from low tension (large compressions) in regions of slow speech to high tension (small compressions) in regions of fast speech. Due to the sign change of β, the opposite is true for time expansion: The audio tension decreases as the relative speaking rate increases, from high

* Pitch resets are perceived as sudden discontinuities in the pitch contour, usually indicating a change of speaker or topic. While the pitch may change drastically from one sentence to the next, if the interphrase pause is naturally long, it is perceived as a continuous variation, instead of as a pitch reset. Only when the pause is artificially shortened is the false percept introduced.

tension (small expansions) in regions of slow speech to low tension (large expansions) in regions of fast speech.

2.3.2 Computing local target compression rates

From audio tension and from a desired global compression (or expansion) rate, we compute local target rates as*

$$r(t) = \max\{1, R_g + (1 - R_g)\gamma(t)\} \quad ,$$

where $r(t)$ is a local target compression or expansion rate and R_g is the desired global compression or expansion rate.

We use these target local compression rates as an input to standard time-scale modification techniques. With synchronous overlap-add (SOLA), for example, we use the local target rates to set, frame by frame, the target offset between the current and previous frames in the output audio signal.

The sequence of local compression (or expansion) rates typically gives overall compression (expansion) rates near the requested global rate, R_g. However, there is no guarantee that this global rate will be achieved. In cases where the global compression rate is important, we add a slow-response feedback loop around the previously described system. This feedback loop acts to correct long-term errors in the overall compression (expansion) rate by adjusting the nominal value of R_g appropriately. The loop's response time must be slow, to avoid distracting artifacts due to rapid changes in the target rate.

3 COMPARISON OF MACH1-COMPRESSED AND LINEARLY COMPRESSED SPEECH

We conducted a listener test comparing Mach1-compressed speech to linearly compressed speech. Parts of this test can be found on our web page: http://www.interval.com/papers/1997–061/.

3.1 Method

Fourteen subjects participated in a listener test to compare comprehension and preferences for Mach1-compressed versus linearly compressed speech. All the subjects were adult professionals, fluent in English and without hearing impairments. None had significant prior experience in listening to time-compressed speech. All aspects of the test, *except* the identity of the compression technique used on each clip, were explained to the subjects before testing.

All the test materials were taken from Kaplan's TOEFL study program [13]. We screened the materials to remove questions based on factual information (e.g., the size of New York). The remaining, uncompressed audio clips range from 111 to 216 wpm.

Each audio clip was sped up twice: once using Mach1 compression and once using linear compression. Since all our audio was single pitched, we used SOLA both as the Mach1-driven compression technique and as the linear-compression technique.

Mach1 compression was done first, with $R_g = 3$, and without global-rate correction (Section 2.3.2). The true compression rate achieved by Mach1 on each audio sample was computed. Each clip was then recompressed linearly to the same global rate that the Mach1 compression achieved. This process gave two versions of each audio sample—one from Mach1, the other from linear compression—both with the same overall compression rate.

The actual compression rates that we achieved using this approach ranged between 2.56 and 4.15 times real time (mean=3.02, median=2.99, s.d.=0.35). The resulting speaking rate ranged from 390 to 673 wpm (mean=500, median=497, s.d.=57).

Each audio sample was assigned to pool A or to pool B. These audio pools were approximately balanced for compression–wpm rates. In the comprehension sections of the test, one half of the subjects heard the pool-A audio clips compressed with Mach1 and the pool-B audio clips compressed linearly. The other one half of the subjects heard the opposite data set: pool-A audio clips compressed linearly and pool-B audio clips compressed with Mach1. Assuming that the two audio pools (and the two subject groups) are well balanced, this technique allows us to use purely within-subjects measures of the difference between Mach1 and linear compression. We tested for differences between the pools as part of our multivariate analysis of variance (MANOVA). No statistically significant difference was found ($F_{1,12} = 0.03$, $p = 0.874$).

In the preference section, both Mach1-compressed and linearly compressed versions of the selected audio samples were played, allowing direct comparisons between the compression techniques.

The test was divided into six sections: Five sections tested comprehensibility of compressed audio clips; the final section tested preference between compression types.

3.1.1 Testing comprehension

Throughout the comprehension sections, each audio clip could be played once only. Each question had to be answered with one of four answers. The choices were displayed after the audio clip finished playing. Response times were measured from the time the answers appeared to the time the subject submitted his answer. The subjects knew that their response times were being measured.

Comprehension was tested in three different categories:

- Short dialog: Three of the comprehension sections were based on 90 short dialogs. Each dialog used one male and one female voice, saying one sentence each. Each audio clip also included one verbal question, spoken by a third voice.
- Long dialog: Another comprehension section was based on 10 long dialogs and 40 questions total. Each dialog used one male and one female voice, each saying 7 to 14 sentences in the course of three to four turns. After the dialog, a sequence of four written questions was shown to the subject.
- Monolog: Another comprehension section was based on eight monologs and 30 questions total. Each monolog used one voice, saying 9 to 15 sentences. After the monolog finished playing, a sequence of three or four written questions was shown to the subject.

3.1.2 Testing subjective preference

The preference section was based on 40 pairs of audio clips. Each pair of clips used either a dialog or a monolog from the previous comprehension sections. Each pair was compressed as described previously. In this section only, the subjects could control freely the playback of each pair of audio clips: They could play either audio clip as often as desired, they could switch back and forth between audio clips, and they could rewind either clip.

Once the subjects listened to at least part of each of the paired audio clips, they could select one as their preference. They were required to make a choice between all pairs of audio. Response times were not measured; instead, the subjects were encouraged to take as much time as they needed to decide between the clips.

3.2 Results

We analyzed the results of the comprehension sections using a two-way, within-subjects MANOVA. The two treatment factors were two compression types (Mach1, linear) and five test sections (three short-dialog sections, one long-dialog section, one monolog section). The mean comprehension rate across all categories was 77%. The mean response time was 7.9 sec. The overall difference in comprehension rates between Mach1-compressed and linearly compressed speech was 17 ± 4percentage points ($F_{1,13} = 69.8$, $p < 0.001$). The overall difference in response times between com-

* In this equation for $r(t)$, we assume that both compression and expansion rates are expressed as numbers greater than 1. The target output frame offset is set to $1/r(t)$ times the input offset for compression, and is set to $r(t)$ times the input offset for expansion.

pression types was –1.2±0.6 sec ($F_{1,13}=17.8$, $p=0.001$). There were also significant interactions between compression type and test section. The differences in comprehension rates between the compression types are shown by section in Table 1. We tested the significance of these differences individually using planned comparisons. The results of those tests are also included in Table 1.*

We also did a regression analysis of the question-by-question comprehension rates (averaged across subjects) versus compression rate. As expected, with linearly compressed speech, comprehension fell with increased compression: $r=-0.25$,† $t_{158}=4.20$, $p<0.001$. In contrast, with Mach1-compressed speech, there was *no* significant comprehension loss with increased compression. Furthermore, the difference in these two comprehension rates increased 34 percentage points with each unit increase in compression rate: $r=0.42$; $t_{158}=5.81$, $p<0.001$. There were no statistically significant correlations between comprehension and speaking rate.

In the preference section, Mach1-compressed speech was chosen 94.8% of the time over linearly compressed speech, for identical global compression rates. This preference rate is clearly different from random selection ($t_{12}=21.9$, $p<0.001$). The Mach1-preference rate increased 10 percentage points with each unit increase in compression rate: $r=0.56$, $t_{38}=4.76$, $p<0.001$. There was no significant correlation between Mach1-preference rate and speaking rate.

4 CONCLUSIONS

Mach1 offers significant improvements in comprehension over linear compression, especially at high compression rates: Comprehension improved by 17 percentage points when Mach1 was used instead of linear compression, at the same global rates. This difference in comprehension increased with increasing compression rate. Listeners preferred Mach1-compressed speech over linearly compressed speech 95% of the time. This preference increased with increasing compression rate.

Short dialogs provided the greatest improvement in comprehension, averaging 23 percentage points. The comprehension improvements were less with the longer clips: 10 percentage points with monologs and 5 percentage points with long dialogs. The large comprehension improvements on short dialogs was due mostly to lowered comprehension of the linearly compressed speech. Since the short dialogs (averaging 23 words) are significantly shorter than the other clips (averaging 144 and 187 words for the long dialogs and monologs), one possible explanation for the lower comprehension of the linearly compressed short dialogs is that the most information is lost at the beginning of the clips, while the subjects adjust to the unnatural speaking style. The absence of a similar decrease in comprehension of Mach1-compressed short dialogs suggests that the listener-adjustment period is much shorter when Mach1 is used.

The comprehension improvements with Mach1 were statistically significant for short dialogs and for monologs. The improvement for long dialogs was not statistically significant. This failure to attain statistical significance may be due to the small test population. Another possibility is there may be confusing interactions between Mach1 and the turn-taking techniques used in conversation. These interactions could have been masked in the short dialogs by the heavily reduced comprehension of linearly compressed speech.

Variable-rate compression of speech is a promising direction for time-scale modification. It should allow us to improve our comprehension rates using approaches suggested by linguistic and text-to-speech studies. It leaves open the question of how best to measure paralinguistic qualities, such as emphasis and relative speaking rate. The Mach1 approach avoids categorical labels and relies on easily measurable acoustic correlates. This approach has proved fruitful, conferring significant improvements in comprehension over linear compression.

* Table 1 uses the Bonferroni t' distribution with $N_c = 5$ [14].

† The correlation-coefficient estimate, r, is tested for significance using $\sqrt{(1-r^2)/(N-2)}$ as its standard error [14].

Table 1: Comprehension-rate differences between Mach1-compressed and linearly compressed speech, by test section. Significance levels for each section are also shown.

Section type	**Comprehension rates** (in percentage points)	
	Average	**Difference (Mach1 – linear)**
short dialogs	70	31.0 ($t'_{52}=8.60, p<0.001$)
	82	14.8 ($t'_{52}=4.13, p<0.001$)
	76	24.3 ($t'_{52}=6.79, p<0.001$)
long dialogs	79	5.4 ($t'_{52}=1.50, p=0.702$)
monologs	81	10.0 ($t'_{52}=2.79, p=0.036$)

ACKNOWLEDGMENTS

We thank Gerald McRoberts and Dan Levitin for their advice on designing the listener test, Jennifer Orton for running the listener tests, Jennifer Smith for her guidance and work in statistical analysis, and Lyn Dupré for her editing.

REFERENCES

[1] P. King, R. Behnke, “The Effect of Time-Compressed Speech on Comprehensive, Interpretive, and Short-Term Listening,” *Human Communication Research,* 15(3): 428–443, 1989.

[2] P. Gade, C. Mills, “Listening Rate and Comprehension as a Function of Preference for and Exposure to Time-Altered Speech,” *Perceptual and Motor Skills,* 68(2): 531–538, 1989.

[3] C. Fulford, “Can Learning be more Efficient? Using Compressed Audio Tapes to Enhance Systematically Designed Text,” *Educational Technology,* 33(2): 51–59, 1993.

[4] J. van Santen, “Assignment of Segmental Duration in Text-to-Speech Synthesis,” *Computer Speech and Language,* 8(2): 95–128, 1994.

[5] M. Withgott, F. Chen, *Computational Models of American Speech,* Center for the Study of Language and Information, 1993.

[6] B. Arons, “Interactively Skimming Recorded Speech,” Ph.D. dissertation, Massachusetts Institute of Technology, 1994.

[7] S. Lee, et al., “Variable Time-Scale Modification of Speech Using Transient Information,” *ICASSP’97,* pp 1319–1322, 1997.

[8] S. Roucous, A. Wilgus, “High Quality Time-Scale Modification for Speech,” *ICASSP’85,* pp 493–496, 1985.

[9] F. Chen, M. Withgott, “The Use of Emphasis to Automatically Summarize a Spoken Discourse,” *ICASSP’92,* pp 229–232, 1992.

[10] L. Stifelman, “The Audio Notebook: Paper and Pen Interaction with Structured Speech,” Ph.D. dissertation, Massachusetts Institute of Technology, 1997.

[11] S. Furui, “On the Role of Spectral Transition for Speech Perception,” *JASA,* 80(4): 1016–1025, 1986.

[12] B. Moore, *Hearing,* Academic Press, 1995.

[13] M. Rymniak, et al., *The Essential Review: Test of English as a Foreign Language,* Kaplan Educational Centers, 1997.

[14] D. Howell, *Statistical Methods for Psychology,* Duxbury Press, 1992.