Original Paper

Phonetica 2009;66:113–126
DOI: 10.1159/000208934

Received: September 10, 2008
Accepted: January 8, 2009

On the Possible Role of Brain Rhythms in Speech Perception: Intelligibility of Time-Compressed Speech with Periodic and Aperiodic Insertions of Silence

Oded Ghitza[a,b] Steven Greenberg[c]

[a]Sensimetrics Corp., Malden, Mass., [b]Boston University, Boston, Mass., and [c]Steven Greenberg, Silicon Speech, Santa Venetia, Calif., USA

Abstract

This study was motivated by the prospective role played by brain rhythms in speech perception. The intelligibility – in terms of word error rate – of natural-sounding, synthetically generated sentences was measured using a paradigm that alters speech-energy rhythm over a range of frequencies. The material comprised 96 semantically unpredictable sentences, each approximately 2 s long (6–8 words per sentence), generated by a high-quality text-to-speech (TTS) synthesis engine. The TTS waveform was time-compressed by a factor of 3, creating a signal with a syllable rhythm three times faster than the original, and whose intelligibility is poor (<50% words correct). A waveform with an *artificial* rhythm was produced by automatically segmenting the time-compressed waveform into consecutive 40-ms fragments, each followed by a *silent* interval. The parameters varied were the length of the silent interval (0–160 ms) and whether the lengths of silence were equal ('periodic') or not ('aperiodic'). The performance curve (word error rate as a function of mean duration of silence) was U-shaped. The lowest word error rate (i.e., highest intelligibility) occurred when the silence was 80 ms long and inserted periodically. This is also the condition for which word error rate increased when the silence was inserted aperiodically. These data are consistent with a model (TEMPO) in which low-frequency brain rhythms affect the ability to decode the speech signal. In TEMPO, optimum intelligibility is achieved when the syllable rhythm is within the range of the high theta-frequency brain rhythms (6–12 Hz), comparable to the rate at which segments and syllables are articulated in conversational speech.

1. Introduction

Speech is an inherently rhythmic phenomenon in which the acoustic signal is transmitted in syllabic 'packets' and temporally structured so that most of the energy

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0031–8388/09/0662–0113
$26.00/0
Accessible online at:
www.karger.com/pho

Oded Ghitza
14 Summer Street, Suite 403
Malden, MA 02148 (USA)
Tel. +1 781 399 0858, ext. 239,
E-Mail oded@sens.com

fluctuations occur in the range between 3 and 20 Hz [e.g., Greenberg, 1999; Greenberg and Arai, 2004]. By using the term 'rhythm' we do not mean that these energy fluctuations are perfectly periodic (in fact, they are not), but rather that there are constraints on syllable duration and energy patterns within and across prosodic phrases. Long syllables are often followed (and preceded) by syllables shorter in duration. And conversely, short syllables are typically preceded and followed by longer ones. Intensity variation follows a similar pattern. Moreover, a slow fluctuation in fundamental frequency (intonation) is also evident in naturally spoken material [e.g., Ladd, 1996; Liberman, 1975]. This rhythmic variation is important for intelligibility and naturalness; speech synthesis studies, for example, have shown that listeners prefer spoken material with a natural, rhythmic structure [e.g., Schroeter, 2008; van Santen et al., 2008]. Does this rhythmic property of speech reflect some fundamental property, one internal to the brain?

In our view, many aspects of spoken language are likely to reflect properties of higher-order cortical processing, not just biomechanical and articulatory constraints. In particular, speech's temporal properties are likely to be constrained not only by how fast the articulators *can* move, but also by how long certain phonetic constituents need to be in order for the signal to be intelligible and sound natural. Although the minimum duration of segments and other phonetic constituents may reflect biomechanical constraints to a certain degree, it is difficult to imagine how biomechanical factors could rigidly specify the length of specific segments and syllables. The suprasegmental properties of speech, especially in view of their variability from language to language, are more likely to be the consequence of factors other than articulation. For example, the range of time intervals (40–4,000 ms) associated with different levels of linguistic abstraction (phonetic feature, segment, syllable, word, metrical foot and prosodic phrase) may reflect temporal constraints associated with neural circuits in the cerebral cortex, thalamus, hippocampus and other regions of the brain. More specifically, certain neural rhythms [e.g., Buszáki, 2006; von Stein and Sarnthein, 2000] could be the reflection of both local and longer-range, transcortical processing. The frequency range over which such rhythms operate (0.5–80 Hz) may serve as the basis for hierarchical synchronization through which the central nervous system processes and integrates sensory information [Freeman, 2007; Lakatos et al., 2005]. It may also reflect a hierarchy of topographic and neural scales, in which higher (and more abstract) levels of processing depend upon information from more extensive cortical areas [von Stein and Sarnthein, 2000].

Such neural rhythms could play an important role in spoken-language comprehension [e.g., Giraud et al., 2007]. A variety of brain-imaging techniques [e.g., Pulvermüller, 1999, for PET and fMRI] allow us to visualize the topography of neural activation associated with linguistic processing in different regions of the cerebral cortex. The specific *timing* of activation can be analyzed using electromagnetic recordings (i.e., magnetoencephalography and electroencephalography). Typically, an increase in oscillatory activity is observed in certain frequency bands, during the performance of specific tasks. Of particular importance are the gamma and theta rhythms [e.g., Bastiaansen et al., 2002; Bastiaansen and Hagoort, 2006; Gevins et al., 1997; Giraud et al., 2007; Luo and Poeppel, 2007]. Theta activity (3–12 Hz) is most closely associated (linguistically) with the syllable (mean duration 200 ms, core range 100–300 ms) [Greenberg, 1999] and the segment (mean duration 80 ms, core range 60–150 ms; [Greenberg et al., 1996], and is thought to

involve some form of sensory-memory comparison process. Gamma oscillations (30–80 Hz)) are most closely associated with neural processing of phonetic constituents and features. Finally, delta oscillations (0.5–3 Hz) may be involved in processing sequences of syllables and words embedded within the metrical foot and prosodic phrase, which could be important for certain aspects of linguistic processing [Roehm et al., 2004].

What is the relation between brain rhythms and spoken language? And why should we expect the perception of speech to be influenced by neural oscillations? Many of the time scales of speech are similar to those of brain rhythms, as mentioned above. Is this temporal similarity a mere coincidence, or does it reflect something deeper? Although this question cannot be answered directly by this (or any other psychophysically based) study, we can begin to investigate this possibility by perturbing the speech input in ways that potentially disrupt the function of brain rhythms and ascertain the impact on intelligibility. In our study, short sentences were interrupted with variable lengths of silence, both periodically and aperiodically. Interruption aperiodicity was used to gauge how underlying neural rhythms interact during the speech-decoding process.

Miller and Licklider [1950] were among the first to systematically examine the temporal parameters associated with the perception of speech. In their study, monosyllabic words were presented over headphones under a variety of signal-processing conditions. Common to all conditions was the use of an analog gating device (a square-wave generator), which interrupted the speech signal over a range of periodic intervals (interruption frequencies ranging between 0.1 and 10,000 Hz). In the simplest set of conditions, the speech signal was gated on for a specific time (e.g., 50 ms) and then gated off for the same interval (i.e., an interruption frequency of 10 Hz in this example). Interruption frequencies above 100 Hz resulted in relatively little degradation (most of which was attributable to spectral distortion). In the range between 1 and 10 Hz, intelligibility declined dramatically. Miller and Licklider speculated that under such conditions intelligibility is governed by the number of 'glimpses' associated with each phonetic segment in the word. When the interruption rate is extremely low, certain segments, syllables or words could not be glimpsed in their entirety, resulting in word-recognition difficulty. We consider this hypothesis further in the 'Discussion' section.

Miller and Licklider also varied the 'speech-time' ratio, the proportion of the gating cycle occupied by the acoustic signal. For example, for a speech-time ratio of 0.25 and an interruption rate of 10 Hz, 25 ms of speech would be followed by 75 ms of silence. Data associated with variation in the speech-time ratio was collected. As the speech ratio increased, so did intelligibility. This result, by itself, is hardly surprising. However, intelligibility became more sensitive to *interruption frequency* when the speech-time ratio was reduced from 50 to 25%. Such a result implies a complex interaction between the specific speech signal glimpsed and the interval of time over which the brain integrates the interleaved speech-silence signal.

There are many issues regarding the temporal processing of speech left unresolved by Miller and Licklider's study. For example, the signal was (for the listening conditions of interest) *interrupted* or *masked* (depending on the condition), which means that some portion of the acoustic signal was discarded (or masked) and hence unheard by the listener. Thus, a certain proportion of the speech information was withheld. Is this speech data loss important for intelligibility, nor not? And does it matter whether the discarded information was in the beginning, middle or end of a syllable or word?

In 1975, Huggins revisited the issue through an ingenious set of experiments. He sought to delineate the underlying temporal factors governing intelligibility. Instead of monosyllabic words, spoken passages (about 150 words long) were used. Listeners continuously 'shadowed' this material (i.e., speaking the words at a comfortable, self-determined delay). Instead of varying the interruption rate, Huggins inserted silence ranging between 16 and 500 ms (the range over which intelligibility was most affected in Miller and Licklider's [1950] study). In contrast to Miller and Licklider's paradigm, no speech was discarded in Huggins' study. Word error rate was measured as a function of speech-time and silence-time durations. With the duration of the speech interval fixed at 63 ms, for example, for small silent gaps (<60 ms), shadowing performance was high. But when the silent gap was long (>150 ms) intelligibility was poor. Huggins suggested that intelligibility depends on 'gap bridging' in these experiments. In his view, there is an ~180-ms-long echoic memory buffer. As long as adjacent speech fragments fall within this buffer interval the brain is able to extract sufficient detail from the acoustic signal to construct a coherent linguistic message. Huggins suggested that the factor governing intelligibility was not phonetic glimpsing per se, but rather some internal time constraint on processing spoken material.

We offer an alternative hypothesis, namely that the decline in intelligibility is the result of a disruption in the syllabic rhythm beyond the limits of what brain neural circuitry can handle. We test this hypothesis by conducting an experiment that extends Huggins' study in a number of important ways.

2. Experiment

One drawback of using semantically plausible material (such as TIMIT or the Harvard-IEEE sentences) in perceptual studies that measure word error rate is the ability of listeners to guess some of the words using contextual information. Semantically unpredictable sentences (SUS) make it more difficult for the listener to decode individual words on the basis of semantic context. For this reason, the materials used in this study are short SUS sentences, taken from the SUSGEN corpus developed by Tim Bunnell (this corpus is used to test the quality and intelligibility of text-to-speech systems).

2.1 SUS Corpus

The experimental corpus used in this study comprised 96 SUS, each approximately 2 s in length (6–8 words per sentence). Each sentence conformed to standard rules of English grammar but was composed of word combinations that are semantically anomalous (e.g., *Where does the cost feel the low night?* and *The vast trade dealt the task*). Text word sequences were generated by Bunnell's SUSGEN program, which produces sentences conforming to a specified grammar and uses a constrained vocabulary [Bunnell et al., 2005].

2.2 Stimulus Preparation

In principle, the speech signals associated with the SUS material could have been produced by a human talker. However, it is difficult to speak such semantically anomalous sentences in a natural

way, particularly in terms of prosody. For this reason, the AT&T Text-to-Speech System (http://www.research.att.com/~ttsweb/tts/demo.php) was used to produce natural-sounding, highly intelligible spoken material with a realistic prosodic rhythm. The AT&T system uses a form of concatenative synthesis (using a high-quality, prerecorded voice) based on unit-selection principles [Schroeter, 2008] and is considered to produce some of the finest quality synthesis of any commercial product [Bunnell, personal commun.]. Each TTS-generated sentence was evaluated carefully for intelligibility and naturalness.

Following synthesis, the waveforms were time-compressed by a factor of 3 using a pitch-synchronous, overlap and add (PSOLA) procedure [Moulines and Charpentier, 1990] incorporated into PRAAT, a speech analysis and modification program (http://www.fon.hum.uva.nl/praat/). In the time-compressed signal, the formant patterns and other spectral properties are altered in duration; however, the fundamental frequency ('pitch') contour remains the same (this is the motivation for using PSOLA methods). Figure 1b shows the time-compressed version of the original, which is shown in figure 1a.

The time-compressed signal served as the baseline waveform for the insertion of silence. First, the waveform was segmented into consecutive 40-ms-long intervals. This segmentation remained fixed throughout. The silences were then inserted; the main parameter was the duration of the silent intervals. The stimulus conditions are summarized in table 1.

To reduce the effect of transients, each speech fragment was multiplied with a 1-ms (rise/fall time) cosine-shaped window. A speech-spectrum-shaped noise was added to the signal after the insertions. The noise level was adjusted to an SNR of 30 dB relative to the power of the signal prior to silence insertions (i.e., condition ×0) in order to perceptually mask the discontinuities associated with the signal processing used to insert silence. This background noise was kept intentionally low in order not to mask speech energy required to decode the signal. Figure 1b–d shows the waveforms for conditions ×0, ×40 and ×80, respectively (see figure caption for details).

The conditions listed in table 1 delineate the periodic class of conditions. An *aperiodic* set was also created, one in which for each condition the silence-time interval was of variable duration, chosen quasi-randomly from one of a set of four intervals equal to 0.5, 0.833, 1.166 and 1.5 of a mean interval equal to the prescribed silence interval indicated in table 1. Any two successive intervals were different from each other.

2.3 Subjects

All 5 listening subjects were young adults (ages 22–27), educated in the US, with normal hearing and no history of auditory pathology. Although the number of listeners is smaller than typical for this type of study, their results (as described in section 2.5) are very consistent with each other.

2.4 Instructions to Subjects

Subjects performed the experiment in their home/office environment using headphones. There were two listening sessions, 'Training' and 'Testing', each lasting approximately 30 min. In the training phase, the subject listened to the original, unprocessed (n = 96) sentences. In the test phase, the subject listened to the same 96 sentences, but this time in processed form. (We believe that the usage of the same sentences for training and for testing is unlikely to have affected the outcome of the experiment. The SUS material is difficult to remember because of the lack of semantic plausibility. To have had a significant impact, listeners would have needed to memorize the specific words played and in their correct order.) The 96 sentences were divided into 12 groups of 8 sentences each, 6 groups for periodic (covering the list of conditions in table 1) and 6 for the aperiodic. Each sentence was listened to only once in the test phase. For each sentence, the subject was instructed to type the words heard (in the order presented) into an electronic file.

The human-subjects protocol for this study was approved by the Institutional Review Board of Sensimetrics Corp.

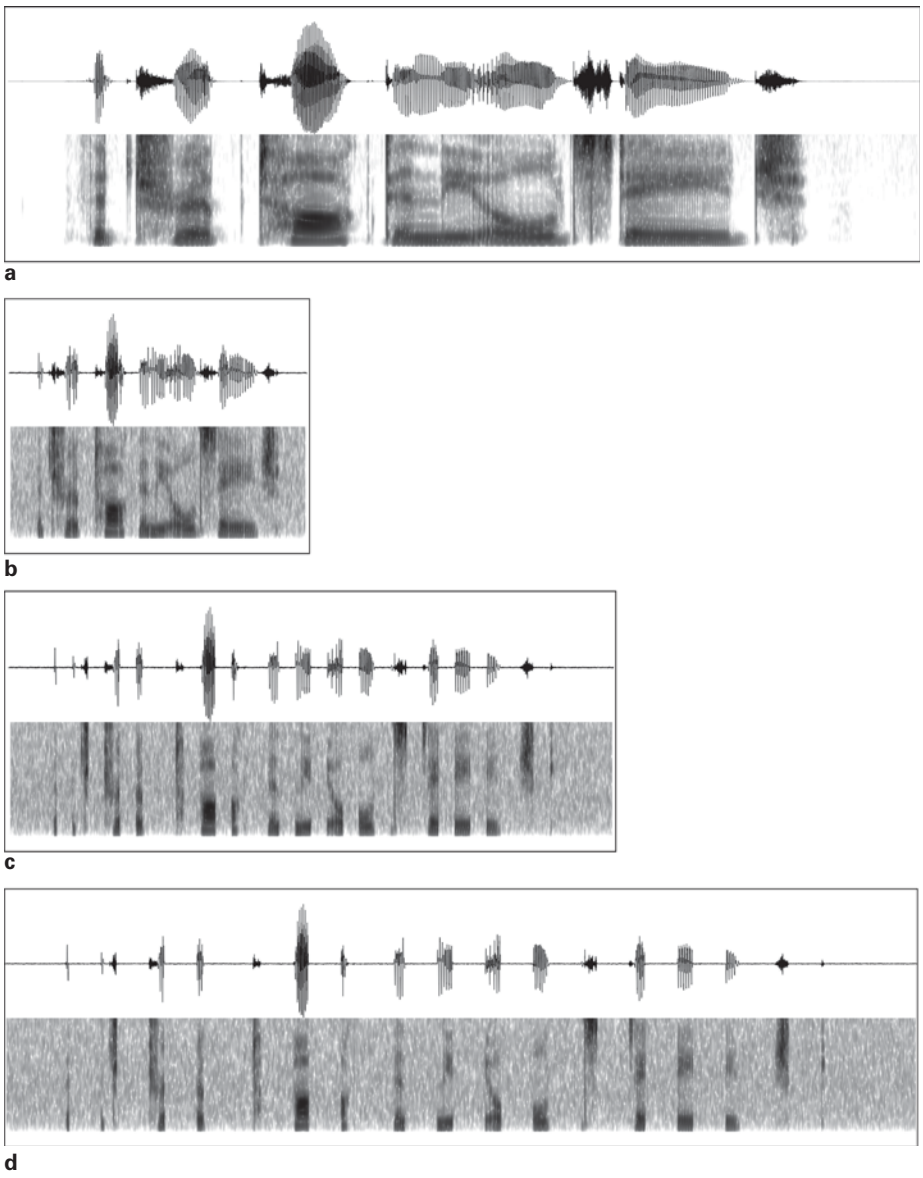

Fig. 1. **a** Waveform (top) and broadband spectrogram (bottom) of the sentence *The trip talked in the old stage*. The waveform duration is 2.4 s, the upper frequency of the spectrogram 5,000 Hz. **b** Same as **a**, time-compressed by a factor of 3. **c** Consecutive 40-ms-long speech intervals of **b**, with 40-ms-long silence insertions. Note that the duration of the processed speech waveform is two thirds the duration of the original signal (i.e., time-compressed by a factor of 1.5 relative to the original). **d** Same as **c** with 80-ms-long silence intervals. The duration of the waveform is the same as the original waveform duration (i.e., no time compression relative to the original). Note that the speech intervals are identical to those in **c**. The background noise visible in the spectrogram was intended to mask discontinuities resulting from inserting silence into the acoustic signal.

Table 1. Experimental conditions used in the study

Condition	Speech interval, ms	Silence interval, ms	Silence/speech	Speed
×0	40	0	0	3
×20	40	20	0.5	2
×40	40	40	1	1.5
×80	40	80	2	1
×120	40	120	3	0.75
×160	40	160	4	0.6

'Speed' refers to the duration of the signal relative to the original, uncompressed sentence.

2.5 Results

2.5.1 Overall

In the training phase, the word error rate was less than 2% for all subjects. Figure 2 shows the mean intelligibility in the testing phase (averaged over the 5 subjects) as a function of the insertion interval. In the absence of insertions (condition ×0), intelligibility is poor (<50% words correct). Intelligibility is equally poor (or worse) when the insertion interval is 160 ms (condition ×160). Interestingly, for insertion intervals between 20 and 120 ms, intelligibility is far better. This is particularly true when the silence interval is 80 ms and inserted periodically. This is also the condition in which there is a significant difference in intelligibility between periodic and aperiodic insertion of silence (the error rate of the latter is nearly twice as high). Two points are noteworthy. First, throughout all conditions, the spectrotemporal information of the speech fragments is time-compressed by a factor of 3. Thus, the U-shape behavior is an unexpected result that is difficult to explain in terms of conventional models of speech perception (see the 'Discussion' that follows). Second, the results indicate a preference for a periodic syllabic rate (particularly for the silence interval condition of 80 ms). Such a result is also difficult to explain with conventional models.

2.5.2 Statistical Analysis

An analysis of variance (ANOVA) was used to compute the statistical significance of the data illustrated in figures 2 and 3. Two factors were used, insertion interval and type of insertion (i.e., periodic vs. aperiodic).

Mauchly's test for sphericity revealed that assumptions of sphericity were not violated. The omnibus repeated-measures two-way (two-variable) ANOVA, with significance level of 0.05, showed that: (1) there is a significant main effect of insertion interval [$F(5, 20) = 24.163$, $p < 0.0001$], (2) there is also a significant main effect of periodicity (i.e., aperiodic vs. periodic) [$F(1, 4) = 16.231$, $p < 0.05$], and (3) there is no significant interaction between insertion interval and type of insertion (periodic/aperiodic) [$F(5, 20) = 2.371$, $p > 0.05$]. It is noteworthy that the ANOVA was not sensitive enough to detect the important 'trend' evident in figure 2, of a significant interaction at the 80-ms insertion interval.

Post-hoc Tukey/Kramer tests revealed that there are significant differences across insertion interval conditions (collapsed across periodicity conditions): (a) the ×0 condition differs from the ×20, ×40, ×80, ×120 conditions, and (b) the ×160 condition differs from the ×20, ×40, ×80, ×120 conditions.

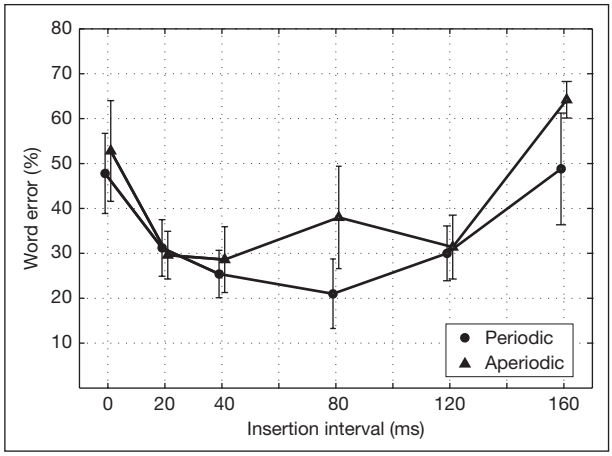

Fig. 2. Intelligibility of time-compressed speech as a function of the duration of inserted silence. Word error rate is plotted as a function of the silence-interval duration (error bars represent the standard deviation of the mean). Speech was time-compressed by a factor of 3. Speech segments are consecutive 40-ms-long intervals and are kept the same for all conditions.

3. Discussion

The data presented in section 2 show that intelligibility of the time-compressed speech is poor (approximately 50% word error) relative to the original (i.e., uncompressed) signal (<2% word error). Insertion of silent intervals markedly improves intelligibility – as long as the silences are between 20 and 120 ms (fig. 2). Conventional models of speech perception have difficulty accounting for this result because they assume a strict decoding of the acoustic signal by the auditory system and higher neural centers.

Can the U-shaped intelligibility curve be explained simply by comparing the temporal properties of the distorted speech with the original, uncompressed waveform? In the 0-ms condition (no silence insertions), there was only a single distortion, that of linear time compression. This compression is sufficient to reduce intelligibility (from the original, unprocessed condition) by 50%. Inserting fragments of silence 20–120 ms in length improves intelligibility dramatically. The best performance is observed when the silence is 80 ms (and inserted periodically). In this condition, the packets of speech information (40-ms intervals of compressed speech) are aligned with the 120-ms intervals of the speech information in the original (uncompressed) speech. In this narrow sense, the acoustic information between the original and compressed version is synchronized. Conditions in which the acoustic alignment between processed and original waveforms is closest would be expected to have the highest intelligibility. Although this is the case, the disparity in intelligibility across the 20- to 120-ms silence insertion conditions is relatively small. Why should an alignment disparity of this magnitude result in so little difference in intelligibility? And why would a comparable difference in acoustic alignment (0-ms and 20-ms, or 120-ms and 160-ms conditions) have such a large impact on decoding the speech signal? Some other factors, quantal-like in nature, are more likely to account for the rather strange pattern of intelligibility. What might they be?

Within the classical framework, intelligibility would not be expected to vary with the length of silence because insertions do not affect the acoustic signal directly, only the temporal distribution of speech information to the auditory system and the

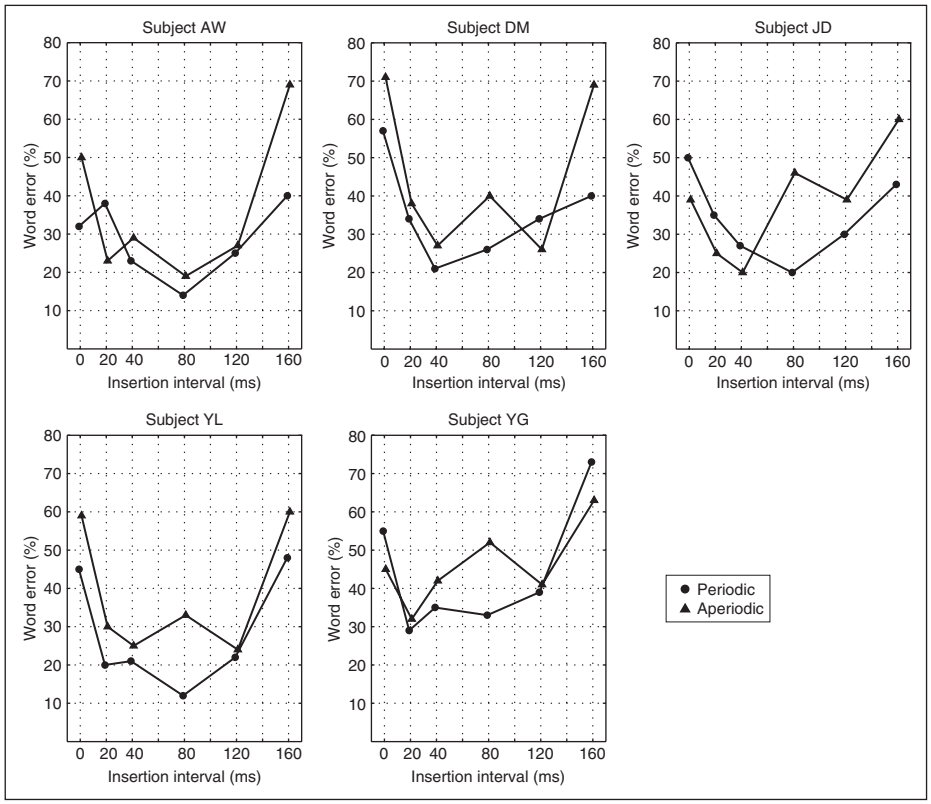

Fig. 3. Same as in figure 2, but the data plotted pertain to each subject individually.

brain. However, it has been known since the studies of Miller and Licklider [1950] and Huggins [1975] (and others) that temporal packaging of the speech signal can exert an enormous impact on intelligibility. Miller and Licklider attributed most of the decline in intelligibility to lost opportunities for glimpsing information in the speech signal. Specifically, they suggested that intelligibility depends on two parameters: (1) the glimpsing rate and (2) the amount of speech information delivered per glimpse. In their experiment, the speech was uncompressed, and the amount of speech information deleted indicated only by the duty cycle. The shorter the duty cycle the smaller the ratio of speech to the overall interruption cycle. For example, for a fixed periodic interruption (i.e., glimpsing rate) of 5 Hz, the word error rate increases from ~30% for a 50% duty cycle to ~70% for a 25% duty cycle.

In the current experiment, glimpsing rate was varied in a different way, namely by changing the duration of the silent interval. The amount of speech information per glimpse was held constant (determined by speech fragment duration – 40 ms – and the time-compression ratio of 3).

Both in our study and in Miller and Licklider's, glimpsing appears to play an important role in intelligibility. However, there is a crucial difference between the two

studies that bears on the neural mechanisms involved in decoding speech. In Miller and Licklider's study, glimpsing rate and speech information per glimpse were confounded – the amount of speech information accessible to listeners was determined, in tandem, by the interaction of interruption rate and duty cycle. Hence, their results could be explained in part by the amount of speech information per glimpse available to the listener. In this sense, Miller and Licklider's study does not directly challenge the conventional models for decoding speech.

In contrast, our study dissociated the factors of glimpsing rate and speech information per glimpse. Because no portion of the acoustic signal was discarded (only time-compressed), and because the speech intervals remained fixed throughout the experiment, the amount of speech information per glimpse was kept constant. Only the glimpsing rate varied, and its rate was directly tied to the length of inserted silence. Any change in intelligibility could therefore be attributed to glimpsing rate per se, rather than to the amount of information contained in the speech signal. This variation in intelligibility is much harder to explain in terms of the standard models of speech perception. It is not just the information in the acoustic signal that is important, but also the timing of the information packets. This is a factor that the standard models do not address.

So, we pose the question: what are the neural mechanisms that distinguish glimpsing rates that are easy to linguistically decode from those that are not? In our view, the answer to this question is that glimpsing rate is governed by endogenous brain rhythms. We assume that neural processes underlying the decoding of speech have oscillations at their core, operating in the gamma, theta and (potentially) delta range. Such neural circuitry may 'prefer' incoming sensory information in the form of temporally constrained packets, compatible for pattern matching and compensating for various forms of background interference (as well as variation in speaking rate). Normally, the operation of these cortical rhythms is 'hidden' from view. Special methods are required to reveal their influence. The aperiodic insertion conditions were designed to reveal the possible role of rhythms.

In Huggins' [1975] study, all of the speech information was preserved. None of it was deleted, similar to what was done in our experiment. Hence, the variation in intelligibility could only have been the result of the temporal distribution of speech information, analogous to what was observed in our own study. However, Huggins did not suggest that his results were the result of brain rhythms. Rather, he suggested that the decline in intelligibility resulted from limitations of short-term working memory. When the silence interval exceeded a certain length, the time between adjacent speech fragments would be too long for the brain to integrate the acoustic signals into a single stream capable of being linguistically decodable. If short-term memory were the primary factor affecting intelligibility, the pattern of intelligibility in our experiments would be very different. First, performance would not necessarily improve as silence is inserted. Why should it, if the major parameter affecting speech decoding is the interval between successive speech fragments? Second, performance would not be different for the ×80 periodic and the ×80 aperiodic conditions since the requirements pertaining to short-term memory are very similar. Nor can the periodic/aperiodic intelligibility differences observed be attributed entirely to the linguistic difficulty of the aperiodically inserted material. If this had been the case, most of the errors would be associated with the same sentences across listeners. Instead, the error distribution is quite variable across listeners and sentences. Although it is possible that short-term

memory does play some role in decoding speech, it is unlikely to be the all-important factor suggested by Huggins.

3.1 Why Is the Intelligibility Curve U-Shaped?

What is particularly striking is the fact that intelligibility *improves* so markedly when the signal is temporally distorted. This demonstrates that intelligibility is not simply a matter of decoding the spectro-temporal pattern – something else is also going on. One possible explanation is that the brain requires a time window with a certain minimum duration to accurately decode the signal, and that the duration of the time window is within a range determined by theta and alpha oscillations. Our results may reflect the degree to which syllabic modulations are matched to these rhythms (internal to the brain). Theta rhythms, in particular, are considered to be involved in some form of communication between distant brain regions [Buszáki, 2006]. The hippocampus, important for short-term memory retrieval, also appears to be involved [Buszáki, 2006].

In order to gain insight into how brain rhythms affect speech decoding we are in the process of developing a phenomenological model called TEMPO. In this model, the process of matching spectro-temporal patterns with phonetic and other types of linguistic elements is temporally controlled (and guided) by *nested* oscillators operating in the delta (<3 Hz), theta (3–12 Hz) and gamma (30–80 Hz) ranges. In speech, the delta range has temporal properties commensurate with phrasal- and lexical-length units (400–4,000 ms), while theta oscillations are associated with individual syllables (100–400 ms) and segments (80–160 ms). Gamma oscillations (>25 Hz) have timing properties potentially relevant to the relatively rapid spectro-temporal (i.e., formant) transitions associated with diphone elements (i.e., consonant-vowel or vowel-consonant, 20–40 ms long). Although all three intervals are important for decoding speech, associated with information at the phonetic, segmental, syllabic, lexical and phrasal levels, the current study focuses on rhythms in the theta range.

The TEMPO model encapsulates this multilevel property of speech, as shown in the block diagram depicted in figure 4. The speech signal is processed by a model of the auditory periphery [e.g., Ghitza et al., 2007], resulting in a spectro-temporal sensory representation. This multichannel information is processed by a Template Matching Circuit (TMC) that matches 'phonetic primitives' to 'memory' neurons by measuring coincidence in firing activity across frequency channels. At this level, time-frequency patterns are matched over relatively short time intervals (approximately 25–50 ms). They are usually formant transitions associated with such phonetic features as place of articulation (important for distinguishing among consonants). In TEMPO, pattern matching at this fast time scale is regulated by gamma oscillations (30–80 Hz).

The neural activity of the phonetic-primitive memory neurons is temporally integrated by a Temporal Sequencing Circuit, with theta oscillations at its core. The theta oscillator operates on a time scale between 50 and 200 ms, and matches syllabic primitives to memory neurons by measuring coincidence in the firing activity of TMC memory neurons across time. The theta oscillations define the time windows in which this temporal integration is performed.

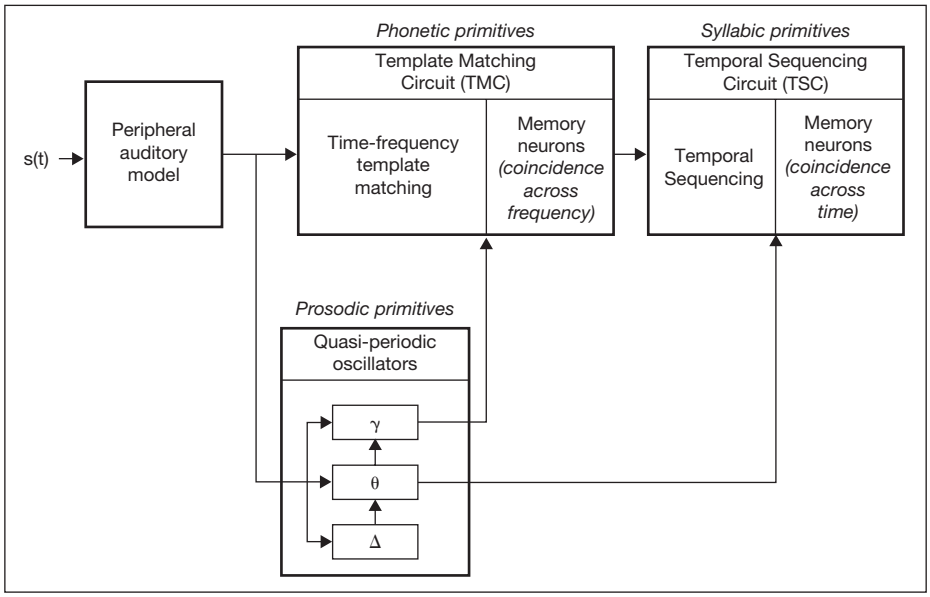

Fig. 4. A block diagram of the TEMPO model. See text for details regarding the model.

3.2 Condition ×80 – Why Does Intelligibility Deteriorate for the Aperiodic Condition?

Although the intelligibility of both the ×80 periodic and ×80 aperiodic conditions is better than the baseline condition (i.e., ×0, time compression with no silence insertions), the aperiodic condition is decoded more poorly than the periodic, and is the only condition where the disparity in intelligibility is large. Moreover, the variability among listeners (as reflected in the standard deviation) is quite high. Recall that the silence intervals in the aperiodic condition ×80 range between 40 and 120 ms (with a mean of 80 ms). Unlike condition ×80, intelligibility in conditions ×40 and ×120 – periodic or aperiodic – is comparable. Moreover, these conditions are decoded far more accurately than the ×80 aperiodic condition. Clearly, there appears to be something about randomly varying the length of the silence in this specific condition that presents problems for decoding speech. (This is the case for all 5 subjects.) What could the reason for this phenomenon be?

One possibility is that the sentences for the ×80 aperiodic condition are more difficult to understand than sentences presented in other conditions. In future studies, the sentential material will be varied in a way to definitively preclude this possibility. However, we believe that sentential difficulty is unlikely to have determined the results because the pattern of errors observed is inconsistent with this possibility. If the relatively poor performance were the result of sentential difficulty, all (or most) of the subjects would have experienced difficulty on precisely the same material and to a comparable degree. However, this is not the case. There is a lot of variability in the specific error patterns associated with this condition, as much as observed for the other conditions.

Rather, we believe that the differential intelligibility is a reflection of the role played by brain rhythms in speech perception: if the decoding process of speech exploits an underlying synchronization mechanism with theta oscillations acting as a pacemaker, then a disruption in the input rhythm is likely to exert a negative impact on intelligibility. By varying the temporal distribution of acoustic information in a quasi-random fashion, it may be that the correspondence between speech input rhythm and the brain rhythms at the core of the decoding process has been disrupted.

There are many parameters associated with cortical rhythms that have not been explored in this study. We have not explored the possibility of interaction between rhythms, nor has the relation between speech fragment length and silence interval been examined. These (and other aspects of brain oscillations) await future experimentation. Rather, our goal has been to ascertain whether brain rhythmicity could play a role in decoding speech. In our view, the answer is affirmative, and therefore provides good reason to continue investigating the role played by brain rhythms in speech perception.

4. Summary

Intelligibility (i.e., word error rate) of spoken sentences was measured as a function of judiciously manipulated changes in syllabic rhythm. The results are surprising and may provide insights into how the speech signal is decoded by the brain. We found that the time-compressed signal (i.e., without insertions of silence) is difficult to understand, and that insertion of silence improves intelligibility, but only over a certain range of durations (20–120 ms). Moreover, the highest intelligibility (which occurs for 80-ms silence insertions) is associated with waveform-energy fluctuations in the theta range (6–12 Hz), which is similar to the syllabic rate of natural speech. We suggest that the poor intelligibility observed in this study may reflect a mismatch between the stimulus energy fluctuations and the frequency range of endogenous neural oscillators that lie at the core of the decoding process of speech. We also found that aperiodic insertions of silence degrade intelligibility, principally for intervals close to 80 ms. This result is at odds with the notion of a short-term memory buffer [hypothesized by Huggins, 1975], but is consistent with a mechanism based on brain rhythms.

In conclusion, this study provides empirical support for the hypothesis that brain rhythms are important in processing and decoding spoken language. Although no hypothesis about internal physiological processes can be fully validated using only psychophysical methods, the perceptual consequences of the acoustic manipulations used in this study suggest a potential role for brain rhythms in speech perception and establishes a behavioral context for future brain-imaging experiments using comparable speech material.

5. Acknowledgments

This study was funded by a research grant from the Air Force Office of Scientific Research. We would like to thank Dr. Willard Larkin for his encouragement and for valuable discussion that led directly to the experiment described in this article. We also thank Prof. Tim Bunnell for providing the sentence list generated by his program for producing semantically unpredictable sentences (SUSGEN). Dr. Ann Syrdal of AT&T provided valuable advice about methods for generating the stimuli used in the study and suggested using the AT&T concatenative synthesis system. Dr. Udi Ghitza provided valuable assistance with the statistical analyses. We would also like to thank Steve Epstein, Klaus Kohler and 2 anonymous reviewers for their valuable suggestions for improving this article. We are also grateful to the individuals who served as listening subjects in the experiment.

References

Bastiaansen, M.; Hagoort, P.: Oscillatory neuronal dynamics during language comprehension. Prog. Brain Res. *159:* 179–196 (2006).

Bastiaansen, M.C.; Berkum, J.J.; Hagoort, P.: Event-related theta power increases in the human EEG during online sentence processing. Neurosci. Lett. *19:* 13–16 (2002).

Bunnell, H.T.; Pennington, C.; Yarrington, D.; Gray, J.: Automatic personal synthetic voice construction. Proc. Interspeech 2005, pp. 89–92.

Buzsáki, G.: Rhythms of the brain (Oxford University Press, New York 2006).

Freeman, W.: My legacy: a launch pad for exploring neocortex. Brain Network Dynamics Conf., Berkeley 2007.

Gevins, A.; Smith, M.E.; McEvoy, L; Yu, D.: High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. Cerebr. Cortex *7:* 374–485 (1997).

Giraud, A.L.; Kleinschmidt, A.; Poeppel, D.; Lund, T.E.; Frackowiak, R.S.J.; Laufs, H.: Endogenous cortical rhythms determine cerebral specialisation for speech perception and production. Neuron *56:* 1127–1134 (2007).

Ghitza, O.; Messing, D.; Delhorne, L.; Braida, L.; Bruckert, E.; Sondhi, M.M.: Towards predicting consonant confusions of degraded speech; in Kollmeier, Klump, Hohmann, Langemann, Mauermann, Uppenkamp, Verhey, Hearing – from sensory processing to perception, pp. 541–550 (Springer, Berlin 2007).

Greenberg, S.: Speaking in shorthand: a syllable-centric perspective for understanding pronunciation variation. Speech Commun. *29:* 159–176 (1999).

Greenberg, S.; Arai, T.: What are the essential cues for understanding spoken language? IEICE Trans. Inf. Syst. *E87:* 1059–1070 (2004).

Greenberg, S.; Hollenback, J.; Ellis, D.: Insights into spoken language gleaned from transcription of the Switchboard corpus. Proc. 4th Int. Conf. Spoken Lang. Proc., 1996, pp. S24–S27.

Huggins, A.W.F.: Temporally segmented speech. Perception Psychophysics *18:* 149–157 (1975).

Ladd, R.: Intonational phonology (Cambridge University Press, Cambridge 1996).

Lakatos, P.; Shah, A.S.; Knuth, K.H.; Ulbert, I.; Karmos, G.; Schroeder, C.E.: An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. J. Neurophysiol. *94:* 1904–1911 (2005).

Liberman, M.: The intonational system of English; PhD thesis, MIT (1975).

Luo H.; Poeppel, D.: Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron *54:* 1001–1010 (2007).

Miller, G.A.; Licklider, J.C.R.: The intelligibility of interrupted speech. J. Acoust. Soc. Am. *22:* 167–173 (1950).

Moulines, E.; Charpentier, F.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. *9:* 453–467 (1990).

Pulvermüller, F.: Words in the brain's language. Behav. Brain Sci. *22:* 253–366 (1999).

Roehm, D.; Schlesewsky, M.; Bornkessel, I.; Frisch, S.; Haider, H.: Fractionating language comprehension via frequency characteristics of the human EEG. Neuroreport *15:* 409–412 (2004).

Santen van, J.P.H.; Mishra, T.; Klabbers, E.: Prosodic processing; in Benesty, Sondhi, Huang, Handbook of Speech Processing, pp. 471–487 (Springer, Berlin 2008).

Schroeter, J.: Basic principles of speech synthesis; in Benesty, Sondhi, Huang, Handbook of speech processing, pp. 413–428 (Springer, Berlin 2008).

Stein von, A.; Sarnthein, J.: Different frequencies for different scales of cortical integration: from local gamma to long range alpha/theta synchronization. Int. J. Psychophysiol. *38:* 301–313 (2000).