Speech Intelligibility is Highly Tolerant of Cross-Channel Spectral Asynchrony

Steven Greenberg[1] and Takayuki Arai[1, 2]

International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704, USA[1]
Department of Electrical and Electronic Engineering, Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo, Japan[2]

Abstract: A detailed auditory analysis of the short-term acoustic spectrum is generally considered essential for understanding spoken language. This assumption is called into question by the results of an experiment in which the spectrum of spoken sentences was partitioned into quarter-octave channels and the onset of each channel shifted in time relative to the others so as to desynchronize spectral information across the frequency plane. Intelligibility of sentential material (as measured in terms of word accuracy) is unaffected by a (maximum) onset jitter of 80 ms or less and remains high (> 75%) even for jitter intervals of 140 ms. Only when the jitter imposed across channels exceeds 200 ms does intelligibility fall below 50%. These results imply that the cues required to understand spoken language are not optimally specified in the short-term spectral domain, but may rather be based on some other set of representational cues such as the modulation spectrogram [S. Greenberg and B. Kingsbury, Proc. IEEE ICASSP, 1997, pp. 1647-1650]. Consistent with this hypothesis is the fact that intelligibility (as a function of onset-jitter interval) is highly correlated with the magnitude of the modulation spectrum between 3 and 6 Hz.

INTRODUCTION

Traditional models of speech recognition (by both human and machine) assume that a detailed auditory analysis of the short-term acoustic spectrum is essential for understanding spoken language. In such models each phonetic segment in the phonemic inventory is associated with a canonical set of (context-dependent) acoustic cues, and it is from such features that phonetic-level constituents are, in principle, identified and placed in sequence to form higher-level linguistic units such as the word and phrase. Significant alteration of these acoustic landmarks should disrupt the decoding process and thereby degrade the intelligibility of speech. We test the validity of this conceptual framework by scrambling the spectro-temporal components of the speech signal beyond all spectrographic recognition [see reference (1 - Figure 1)] and demonstrate that spectral desynchronization of up to 140 ms has relatively little impact on intelligibility. Analysis of the spectrally desynchronized waveforms in terms of the low-frequency (3-6 Hz) modulation spectrum indicates that this alternative representation provides an effective means to predict intelligibility over a wide range of spectrally asynchronous conditions reminiscent of acoustic reverberation. Under optimal listening conditions the modulation spectral information germane to intelligibility is concentrated in the lower frequency channels (< 1.5 kHz). These same channels appear to play a considerably less unimportant role in processing speech under conditions of significant spectral asynchrony, ceding their dominance to frequencies above 1.5 kHz. This differential capability may underlie the robust nature of spoken language and provide a principled basis for understanding the nature of linguistic deficit sustained by the hearing impaired under deleterious acoustic conditions characteristic of the real world.

EXPERIMENTAL METHODS

The spectrum of spoken sentences (sampled at 16 kHz, with 16-bit resolution, and derived from the TIMIT corpus) was partitioned into 19 channels. The lowest channel encompassed all energy below 265 Hz, while the other 18 channels partitioned the remainder of the spectral domain (265-6000 Hz) into quarter-octave intervals. The output of each channel was shifted in time relative to the baseline in such a manner as to approximate a uniform distribution of temporal intervals ranging from 0 to a maximum delay, D_{max}, where D_{max} varied between 60 and 240 ms (in steps of 20 ms). The delay between adjacent channels was constrained so as to exceed one quarter of the maximum delay (i.e., $D_{max}/4$) in order to preclude the generation of local pockets of high temporal correlation. The sampling procedure was adapted to insure that the distribution was uniform in the presence of such local decorrelation, making it relatively straightforward to characterize the gross statistical properties of the delay patterns. It is thus possible to estimate both the mean and median of the distribution ($D_{max}/2$) as well as the range of delays spanned by a specified proportion of the distribution. The effect of this asynchrony procedure is to "jitter" the spectral information relative to the original in a fashion reminiscent of reverberation.

Such spectrally jittered sentences were digitally presented at a comfortable listening level over headphones to 30 individuals, all of whom were native-speakers of American English with no known history of hearing impairment. Each subject listened to 40 sentences spoken by different individuals. Two different delay patterns were used for each sentence in order to minimize the impact of a specific asynchrony pattern on the intelligibility patterns. In order to minimize potential learning effects, each sentence was presented under two different conditions, one with a relatively large degree of asynchrony (160-240 ms maximum delay), the second with a relatively small degree of asynchrony (60-140 ms). Subjects listened to a total of 80 sentences, each of which could be repeated up to (a maximum of) four times. A brief practice session preceded collection of the intelligibility data. The listener was instructed to type the words heard (in their order of occurrence) into a Sun workstation. The intelligibility score for each sentence was computed by dividing the number of words typed correctly by the total number of words in the spoken sentence. Errors of omission, insertion and substitution were not taken into account in computing this percent-correct score. Speech intelligibility was computed across subjects and sentences for each number of words in the spoken sentence. Because there was no significant difference in intelligibility between the two asynchrony patterns used, performance data were pooled across these conditions. The experimental methods are described in greater detail in (1).

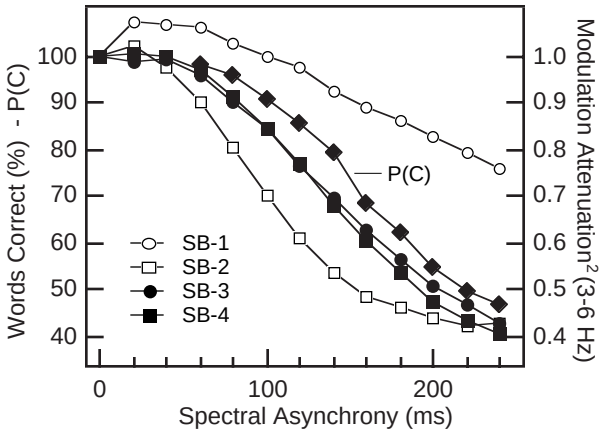

FIGURE 1. Speech intelligibility, P(C), averaged across listeners (30), sentence conditions (40) and asynchrony patterns (2), plotted alongside the square of the attenuation of the *normalized* modulation spectral index in the 3-6 Hz region for four separate sub-bands.

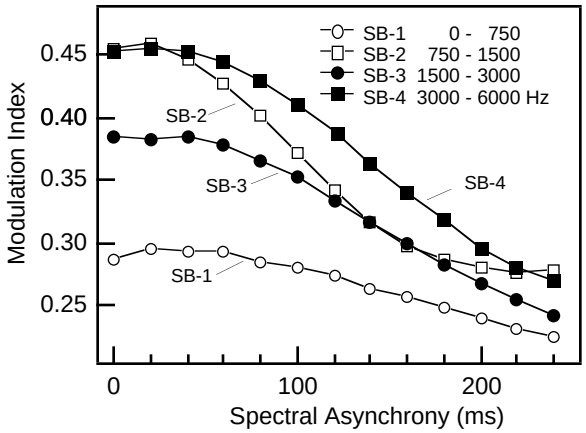

FIGURE 2. Attenuation pattern of the *unnormalized* modulation spectral power in the 3-6 Hz region for four separate sub-bands (SB). The boundaries of these sub-bands are co-terminous with the quarter-octave channels used to spectrally desynchronize the signal. Thus, each of the upper three sub-bands spans four channels, while the lowest contains seven.

INTELLIGIBILITY'S RELATION TO THE MODULATION SPECTRUM

Speech intelligibility, P(C), as a function of (maximal) spectral asynchrony is illustrated in Figure 1. Although intelligibility progressively declines as the degree of spectral asynchrony increases, it is of interest that word accuracy exceeds 75% for the 140-ms asynchrony condition despite the fact that the *average* magnitude of spectral asynchrony approaches the mean duration of a phonetic segment [72 ms - cf. (3)] in this subset of the TIMIT corpus. Even when the asynchrony approaches 200 ms, intelligibility is ca. 50%. These results indicate that linguistically relevant information can be extracted from the speech signal even when the pattern of spectral asynchrony spans two or more phonetic segments. Such intelligibility data are difficult to reconcile with spectral, phone-based models [e.g., (6)] of (human) speech recognition.

An alternative means of representing linguistically relevant information in the speech signal is provided by the modulation spectrum. This representation quantifies the low-frequency (< 12 Hz) acoustic modulation pattern associated with movement of the lips, jaw and tongue during production. The intelligibility of speech vitally depends on the integrity of the modulation spectrum between 3 and 6 Hz under highly reverberant conditions (5). Attenuation of the power in this spectral region via low-pass filtering is known to deleteriously affect the ability to understand spoken language (2, 4). It is therefore of interest to ascertain whether the intelligibility of the TIMIT-sentence material bears a systematic relationship to the low-frequency power of the modulation spectrum. Towards this end, the modulation spectrum between 1 and 12 Hz was computed for each of four spectral sub-bands. The lowest sub-band (SB-1) encompassed the region below 750 Hz, while each of the remaining sub-bands spanned a range of an octave. Spectral desynchronization reduces the power in the modulation spectrum across all four sub-bands (Figures 1 and 2). However, the magnitude of the *average* attenuation (across sentences) in the 3-6 Hz region is considerably smaller for the lowest sub-band (Figure 1) which exhibits significantly less power in this region of the modulation spectrum than the others (Figure 2). The normalized attenuation pattern of the low-frequency modulation spectrum suggests that the decline in intelligibility is most highly correlated with the modulation characteristics of sub-bands 3 and 4 (1.5 - 6 kHz), particularly at intermediate-to-long intervals of spectral asynchrony. A linear, piece-wise analysis of these data reveal a dramatic change in the perceptual weight accorded the lowest and highest sub-bands as the spectral asynchrony increases [cf. (1 - Figure 4)]. For small degrees of asynchrony (≤ 100 ms) the intelligibility data are most highly correlated with the modulation spectral characteristics of the lowest sub-band. At moderate degrees of asynchrony (ca. 120 ms) the modulation spectral properties of sub-band 3 become dominant which, in turn, are superseded in importance by those of the highest sub-band (3-6 kHz) at longer asynchronies.

REFERENCES

1. Arai, T. and Greenberg, S. "Speech intelligibility in the presence of cross-channel spectral asynchrony," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, 1998, in press.
2. Drullman, R., Festen, J. M. and Plomp, R. *J. Acoust. Soc. Am.* **95**, 1053-1064 (1994).
3. Greenberg, S., Hollenback, J. and Ellis, D. "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus." *International Conference on Spoken Language Processing*, Philadelphia, 1996, pp. S32-35.
4. Greenberg, S. and Shire, M. "Temporal factors in speech perception," in *CSRE-based Teaching Modules for Courses in Speech and Hearing Sciences*, London, Ontario: AVAAZ Innovations, 1997, pp. 91-106.
5. Houtgast, T. and Steeneken, H. *J. Acoust. Soc. Am* **77**, 1069-1077 (1985).
6. Klatt, D. H. *J. Phonetics* **7**, 279-312 (1979).