SPEECH INTELLIGIBILITY IN THE PRESENCE OF CROSS-CHANNEL SPECTRAL ASYNCHRONY

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

International Computer Science Institute[1]
1947 Center Street, Berkeley, CA 94704, USA

Department of Electrical and Electronic Engineering[2]
Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo, Japan

ABSTRACT

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 axis. Human listeners are remarkably tolerant of cross-channel spectral asynchrony induced in this fashion. Speech intelligibility remains relatively unimpaired until the average asynchrony spans three or more phonetic segments. Such perceptual robustness is correlated with the magnitude of the low-frequency (3-6 Hz) modulation spectrum and thus highlights the importance of syllabic segmentation and analysis for robust processing of spoken language. High-frequency channels (>1.5 kHz) play a particularly important role when the spectral asynchrony is sufficiently large as to significantly reduce the power in the low-frequency modulation spectrum (analogous to acoustic reverberation) and may thereby account for the deterioration of speech intelligibility among the hearing impaired under conditions of acoustic interference (such as background noise and reverberation) characteristic of the real world.

1. 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 (e.g., [11] [13]). 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. These features are viewed as the representational gateway through which to infer the composition and sequence of articulatory gestures associated with the signal's production [12] [15] [16].

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 and demonstrate that spectral desynchronization 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 relatively unimportant role in processing speech under conditions of significant spectral asynchrony, ceding their dominance to sub-bands encompassing 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.

2. EXPERIMENTAL METHODS

2.1 Signal Processing of Sentential Material

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 (Figure 1) in a fashion reminiscent of reverberation.

2.2 Stimulus Presentation

Such spectrally jittered sentences were digitally presented (16 kHz sample rate, 16-bit resolution) at a comfortable listening level (adjusted by the subject) over headphones (Bang and Olufson #F-2) 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 (balanced across sentence length and lexico-syntactic complexity), spoken by different speakers (equally divided by gender, but all speaking a relatively homogeneous, Western U.S. dialect). 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 (5 sentences) preceded collection of the intelligibility data.

2.2 Data Collection and Analysis

The listener was instructed to type the words heard (in their order of occurrence) into a Sun computer. 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

IEEE International Conference on Acoustics, Speech and Signal Processing, Seatle, pp. 933-936, 1998

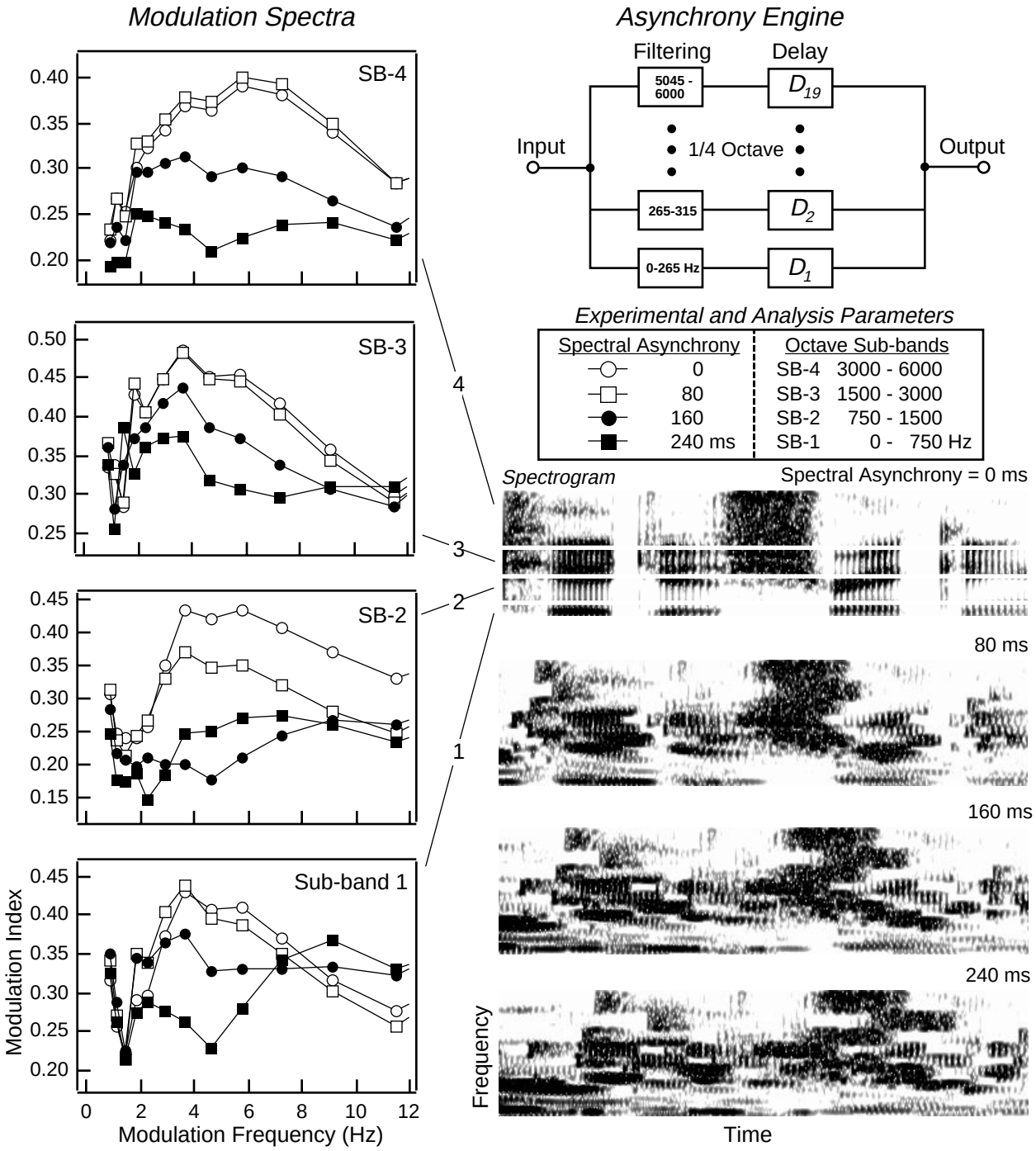

Figure 1. Stimulus generation procedure (upper right) derived from desynchronization of 19 quarter-octave channels. Spectral asynchronies ranged between 60 and 240 ms (in 20-ms steps). The spectrographic representation of a single sentence is shown in the lower right for three degrees of desynchronization (80, 160 and 240 ms) and compared with the spectrogram of the same sentence produced with no asynchrony. The modulation spectra of the sentence material were computed for each of four separate sub-bands (of octave bandwidth, save for the lowest). Representative modulation spectra are shown for the sentence displayed, partitioned among sub-bands and degree of spectral asynchrony. There is a significant attenuation of the modulation spectrum in the range between 3 and 6 Hz for the higher degrees of spectral asynchrony, analogous to the effects exerted on the modulation spectrum by acoustic reverberation.

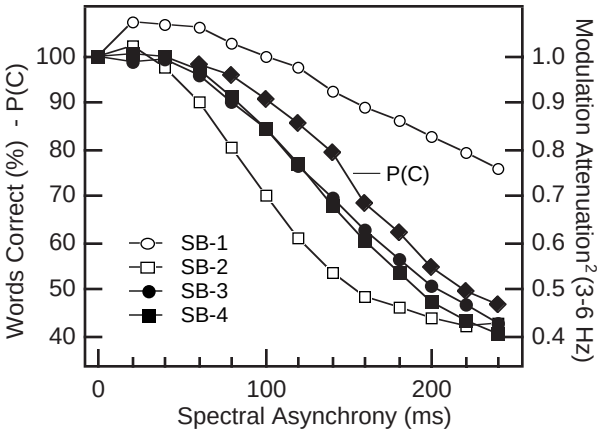

Figure 2. Speech intelligibility, P(C), averaged across listeners (30), sentence conditions (40) and asynchrony patterns (2), plotted alongside the square of the attenuation of the *normalized* modulation spectral index in the 3-6 Hz region for four separate sub-bands.

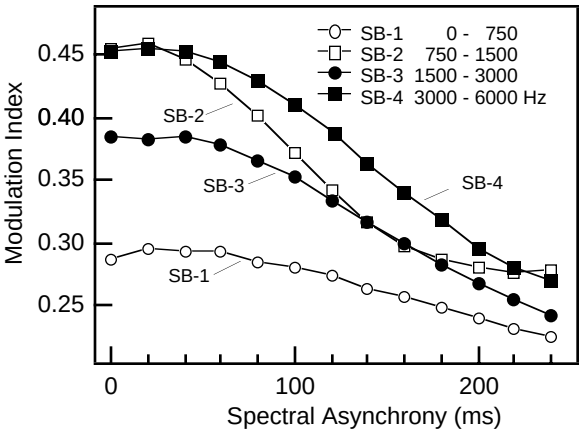

Figure 3. Attenuation pattern of the *unnormalized* modulation spectral power in the 3-6 Hz region for four separate sub-bands (SB). The boundaries of these sub-bands are co-terminous with the quarter-octave channels used to spectrally desynchronize the signal. Thus, each of the upper three sub-bands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

spoken sentence. Because there was no significant difference in intelligibility between the two asynchrony patterns used, performance data were pooled across these conditions.

3. SPEECH INTELLIGIBILITY AS A FUNCTION OF SPECTRAL ASYNCHRONY

Speech intelligibility, P(C), as a function of (maximal) spectral asynchrony is illustrated in Figure 2. 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) in this subset of the TIMIT corpus. Even when the asynchrony exceeds 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., [11] [13]) of (human) speech recognition.

4. INTELLIGIBILITY'S RELATION TO THE MODULATION SPECTRUM

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 [9]. Attenuation of the power in this spectral region via low-pass filtering is known to deleteriously affect the ability to understand spoken language [1] [3] [8]. 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 (Figure 1). However the magnitude of the *average* attenuation (across sentences) in the 3-6 Hz region is considerably smaller for the lowest sub-band (Figures 2 and 3) which exhibits significantly less power in this region of the modulation spectrum than the others (Figure 3).

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 (Figure 4). For small degrees of asynchrony (≤100 ms) the intelligibility data are most closely 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.

5. IMPLICATIONS FOR MODELS OF SPEECH RECOGNITION

Spectral desynchronization of the acoustic waveform imposes certain alterations on the speech signal similar to those associated with reverberation. In both instances there is a frequency-selective temporal smearing that has the effect of "blurring" syllabic boundaries in the waveform and in reducing the amount of power in the 3-6 Hz region of the modulation spectrum [2] (Figure 1). Such conditions generally interfere with the understanding of spoken language and imply that recognition vitally depends on the ability to parse and decode the speech signal into syllabic entities (whose duration typically span 150-300 ms [7]).

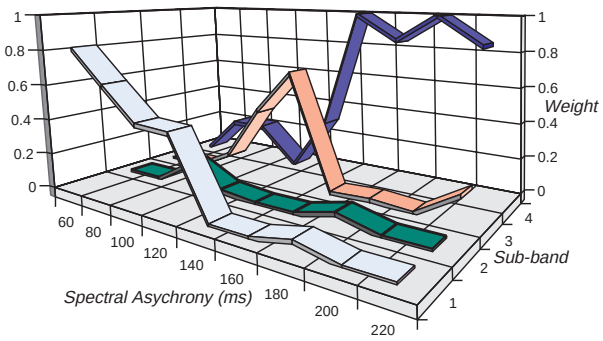

Figure 4. Perceptual weighting functions for each sub-band across asynchrony conditions. Sub-band 1 is most highly weighted for small amounts of spectral asynchrony while sub-band 4 is most highly weighted for large amounts of asynchrony. Weighting functions are derived from data in Figure 2.

The hearing impaired typically manifest little (if any) deficit in intelligibility under high S/N and non-reverberant conditions. Placed in an environment with considerable background noise or reverberation their ability to understand speech rapidly deteriorates. Because much of the energy in the speech signal lies below the region of the spectrum most prominently affected by hearing impairment (>3 kHz), conventional models (such as the Articulation Index (AI) [4] [10] or Speech Transmission Index (STI) [10]) have experienced considerable difficulty predicting speech intelligibility from a *unitary* measure of spectral sensitivity [14]. In quiet, the most highly predictive auditory parameter of intelligibility is the pure-tone threshold *below* 2 kHz, while in noise and reverberation the single best predictor is auditory sensitivity *above* 2 kHz [14].

The current results provide a potential resolution of this seeming paradox, for they imply that the perceptual weight associated with a given frequency region dynamically adapts to the demands of the acoustic environment. In non-reverberant, high S/N conditions the low-frequency channels appear to play an important, if not dominant, role in decoding the speech signal. The relatively small amount of modulation spectral power in the lower frequency channels (Figure 3) implies that the form of linguistic decoding undertaken in this lowest sub-band may focus on relatively short-term properties of the signal, commensurate with traditional models of phonetic feature extraction [11]. The high-frequency (>3 kHz) channels appear to play a particularly important role in syllable segmentation [5], a capacity of potential significance under noisy and reverberant conditions where the ability to extract finer-grained phonetic information is especially compromised. The robustness of spoken language may thus lie in its multifarious capacity for encoding linguistic information across a wide span of time scales [6] and frequency regions.

6. ACKNOWLEDGMENTS

This research was funded, in part, by grants from the U.S. Department of Defense (MDA 904-94-C6196) and the National Science Foundation (SBR-9720398). We thank Jim Matisoff, John Ohala and Tom Shannon for their assistance in recruiting experimental subjects, as well as those students at the University of California, Berkeley who willingly gave of their time (and ears). We also thank Hynek Hermansky, Noboru Kanedera and Nelson Morgan for valuable discussion germane to this paper.

7. REFERENCES

[1] Arai, T., Hermansky, H. Pavel, M. and Avendano, C. "Intelligibility of speech with filtered time trajectories of spectral envelopes," *International Conference on Spoken Language Processing*, Philadelphia, pp. 2490-2493, 1996.

[2] Avendano, C. and Hermansky, H. "Study on the dereverberation of speech based on temporal envelope filtering." *International Conference on Spoken Language Processing*, Philadelphia, 1996, pp. 889-992.

[3] Drullman, R., Festen, J. M. and Plomp, R. "Effect of temporal envelope smearing on speech reception." *J. Acoust. Soc. Am.*, 95: 1053-1064, 1994.

[4] Dubno, J. R. and Dirks, D. D. and Schaefer, A. B. "Stop-consonant recognition for normal-hearing listeners and listeners with high-frequency hearing loss: II. Articulation index predictions." *J. Acoust. Soc. Am.*, 85: 355-364, 1989.

[5] Grant, K. W. and Walden, B. E. "Spectral distribution of prosodic information." *J. Speech Hearing Res.*, 39: 228-238, 1996.

[6] Greenberg, S. "On the origins of speech intelligibility in the real world." ESCA Workshop on *Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, 1996, pp. 23-32.

[7] 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.

[8] 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.

[9] Houtgast, T. and Steeneken, H. "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria." *J. Acoust. Soc. Am*, 77: 1069-1077, 1985.

[10] Humes, L. E., Dirks, D. D., Bell, T.S. and Ahlstrom, C. "Application of the Articulation Index and the Speech Transmission Index to the recognition of speech by normal-hearing and hearing-impaired listeners." *J. Speech Hear. Res.*, 29: 447-462, 1986.

[11] Klatt, D. H. "Speech perception: A model of acoustic-phonetic analysis and lexical access." *J. Phonetics*, 7: 279-312, 1979.

[12] Liberman, A. M. and Mattingly, I. G. "A specialization for speech perception." *Science*, 243: 489-494, 1989.

[13] Pisoni, D. B. and Luce, P. A. "Acoustic-phonetic representations in word recognition," in *Spoken Word Recognition* U.H. Frauenfelder and L. K. Tyler (Eds.), MIT Press: Cambridge, 1987, pp. 21-52.

[14] Smoorenburg, G. F. "Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram." *J. Acoust. Soc. Am.*, 91: 421-437, 1992.

[15] Stevens, K. N. "On the quantal nature of speech." *J. Phonetics* ,14: 373-382, 1989.

[16] Stevens, K. N. "Applying phonetic knowledge to lexical access," *Eurospeech*, Madrid, 1995, pp. 3-11.