Proc. ICSLP 98

SPEECH INTELLIGIBILITY DERIVED FROM EXCEEDINGLY SPARSE SPECTRAL INFORMATION

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

International Computer Science Institute[1]
1947 Center Street, Berkeley, CA 94704, USA
[steveng, rosaria@icsi.berkeley.edu]

Department of Electrical and Electronics Engineering[2]
Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo, Japan
[arai@sophia.ac.jp]

ABSTRACT

Traditional models of speech assume that a detailed auditory analysis of the short-term acoustic spectrum is essential for understanding spoken language. The validity of this assumption was tested by partitioning the spectrum of spoken sentences into 1/3-octave channels ("slits") and measuring the intelligibility associated with each channel presented alone and in concert with the others. Four spectral channels, distributed over the speech-audio range (0.3-6 kHz) are sufficient for human listeners to decode sentential material with nearly 90% accuracy although more than 70% of the spectrum is missing. Word recognition often remains relatively high (60-83%) when just two or three channels are presented concurrently, despite the fact that the intelligibility of these same slits, presented in isolation, is less than 9% (Figure 2). Such data suggest that the intelligibility of spoken language is derived from a compound "image" of the *modulation* spectrum distributed across the *frequency* spectrum (Figures 1 and 3). Because intelligibility seriously degrades when slits are desynchronized by more than 25 ms (Figure 4) this compound image is probably derived from *both* the amplitude and phase components of the modulation spectrum, and implies that listeners' sensitivity to the modulation phase is generally "masked" by the redundancy contained in full-spectrum speech (Figure 5).

1. INTRODUCTION

Classical models of speech recognition (by both human and machine) assume that a detailed analysis of the short-term acoustic spectrum is required for understanding spoken language (e.g., [9] [11]). In such models, each phonetic segment in the phonemic inventory is associated with a canonical set of 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 reducing the spectral cues to a bare skeleton of their normal representation and measuring the intelligibility of sentential material processed in this fashion (Experiment 1). The intelligibility of such sparse spectral signals is far higher than would be predicted by such spectrally formulated frameworks as the Articulation Index [7], and suggests that many of the canonical spectro-temporal cues of phonetic features may not be truly essential for understanding spoken language (at least under optimum listening conditions), as long as the modulation pattern distributed across the frequency spectrum incorporates certain properties of the original, unfiltered signal (cf. Figure 1).

It has been proposed that the intelligibility of speech crucially depends on the integrity of the modulation spectrum's amplitude in the region between 3 and 8 Hz [1] [2] [3] [4] [6]. Experiment 2 tests the validity of this premise by imposing a systematic time-delay pattern on the 4-slit compound and measuring the impact on word intelligibility. Asynchronies as short as 50 ms result in a precipitous decline in intelligibility, demonstrating the importance not only of the amplitude spectrum of the modulation waveform, but also its phase pattern for decoding spoken language.

2. EXPERIMENTAL METHODS

2.1 Signal Processing of Sentential Material

Stimuli were read sentences derived from the TIMIT corpus (spoken by male and female speakers in roughly equal measure, and spanning all major dialect regions of American English). The signals were sampled at 16 kHz and quantized with 16-bit resolution.

Each sentence was spectrally partitioned into 14 1/3-octave-wide channels (using an FIR filter whose slopes exceeded 100 dB/octave) and the stimulus for any single presentation consisted of between 1 and 4 channels presented concurrently. The passband of the lowest-frequency slit was 298-375 Hz, that of the second lowest, 750-945 Hz, that of the third, 1890-2381 Hz, while the passband of the highest-frequency slit was 4762-6000 Hz. Adjacent slits were separated by an octave in order to minimize intermodulation distortion and masking effects potentially arising from the interaction of non-continuous, spectrally proximal components. The spectrographic representation and associated waveform for each slit is illustrated in Figure 1.

2.2 Stimulus Presentation

Such spectrally partitioned sentences were presented at a comfortable listening level (adjusted by the subject) over high-quality headphones to individuals situated in a sound-attenuated room. All were native speakers of American English with no known history of hearing impairment. Each subject listened to between 130 and 136 different sentences (depending on the experiment - see Figures 2 and 4), each of which could be repeated up to four times. A brief practice session (5 sentences) preceded collection of the experimental data. Subjects were financially remunerated for their time and effort.

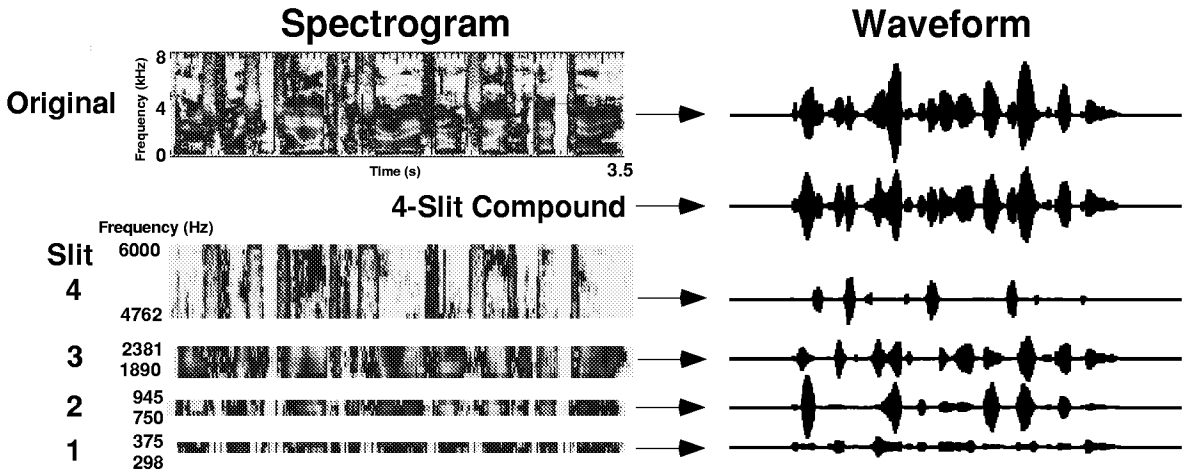

Figure 1. Spectrographic and time-domain representations of a representative sentence ("The most recent geological survey found seismic activity") used in the current study. The slit waveforms are plotted on the same amplitude scale, while the scale of the original, unfiltered signal is compressed by a factor of five. The frequency axis of the spectrographic display of the slits has been non-linearly expanded for illustrative brevity. Note the quasi-orthogonal temporal registration of the waveform modulation pattern across frequency channels. The potential significance of this pattern is discussed in Section 5.

2.3 Data Collection and Analysis

Each listener was instructed to type the words heard (in their order of occurrence) into a computer. The intelligibility score for each sentence was computed by dividing the number of words typed correctly (misspellings notwithstanding) 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. Intelligibility data were pooled across sentence and speaker conditions for each listener. The variance across subjects was on the order of 1-2%, enabling the data to be pooled across listeners as well. 13 listeners participated in Experiment 1 and a separate set of 17 individuals performed Experiment 2.

3. SPEECH INTELLIGIBILITY DERIVED FROM SPECTRAL SLITS

The speech intelligibility associated with each of the fifteen slit combinations is illustrated in Figure 2. Four slits, presented concurrently, result in nearly (but not quite) perfect intelligibility (89%), providing a crucial reference point with which to compare the remaining combinations. A single slit, played in the absence of other spectral information, results in poor intelligibility (2-9%). The addition of a second slit increases word accuracy, but intelligibility is highly dependent on both spectral locus and channel proximity. The two center channels (slits 2 and 3) are associated with the highest degree of intelligibility (60.4%), while the most spectrally distant slits (1 and 4) are only slightly more intelligible than either channel alone (13% combined, versus 2% and 4% separately). However, relative spectral proximity does not always result in higher intelligibility (compare slits 1+2 [28.6%] and 3+4 [30.2%] with slits 1+3 [29.8%] and slits 2+4 [37.1%]). The addition of a third slit results in significantly higher word accuracy, particularly when both slits 2 and 3 are present (78.2 and 82.6%). The omission of either slit results in a much lower level of intelligibility (47.1 and 51.7%). Clearly, some property of the mid-frequency region (750 - 2381 Hz) is of supreme importance for intelligibility (cf. [10] for a complementary perspective on this issue). The current data are in accord with the results reported in [12], although the level of intelligibility obtained by Warren and colleagues (up to 70% for a single slit centered ca. 1500 Hz) far exceeds that obtained in the current study.

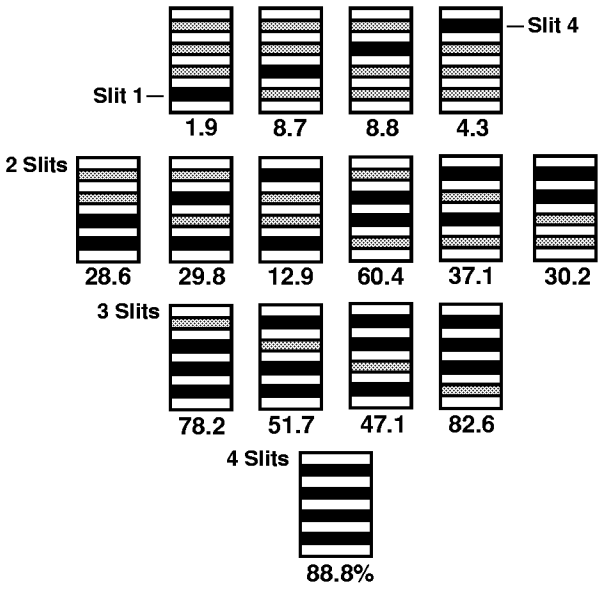

Figure 2. Intelligibility of spectral-slit sentences under 15 separate listening conditions. Baseline word accuracy is 88.8% (4-slit condition). The intelligibility of the multiple-slit signals is far greater than would be predicted on the basis of word accuracy (or error) for individual slits presented alone. The region between 750 and 2400 Hz (slits 2 and 3) provides the most important intelligibility information.

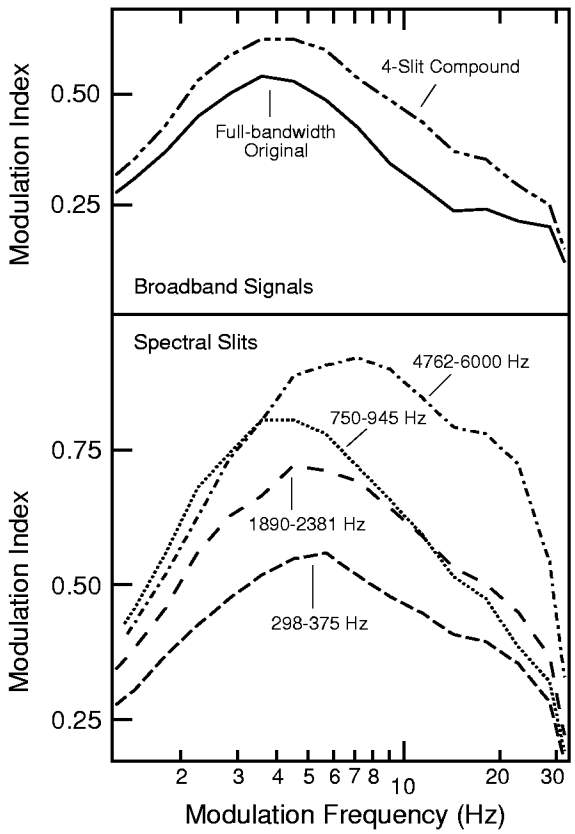

Figure 3. The modulation spectrum (amplitude component) associated with each 1/3-octave slit, as computed for all 130 sentences presented in Experiment 1 [bottom panel]. The peak of the spectrum (in all but the highest channel) lies between 4 and 6 Hz. Its magnitude is considerably diminished in the lowest frequency slit. Also note the large amount of energy in the higher modulation frequencies associated with the highest frequency channel. The modulation spectra of the 4-slit compound and the original, unfiltered signal are illustrated for comparison [top panel].

4. MODULATION SPECTRA OF THE SPECTRAL SLITS

The modulation spectra associated with each of the spectral slits are illustrated in Figure 3. The modulation spectral contours of the three lower slits are similar in shape, all exhibiting a peak between 4 and 6 Hz, consistent with the modulation spectrum of spontaneous speech [5]. The uppermost slit possesses significantly greater energy in the region greater than 5 Hz, reflecting the sharp onset and decay characteristics of this channel's waveform envelope (Figure 1).

The modulation spectrum of the full-band, unfiltered signal is similar in contour to that of the four-slit compound (Figure 3) although it is slightly lower in magnitude (as measured in terms of the modulation index). The similarity of these modulation spectra is consistent with their high correlation of intelligibility (unfiltered sentences completely [100%]

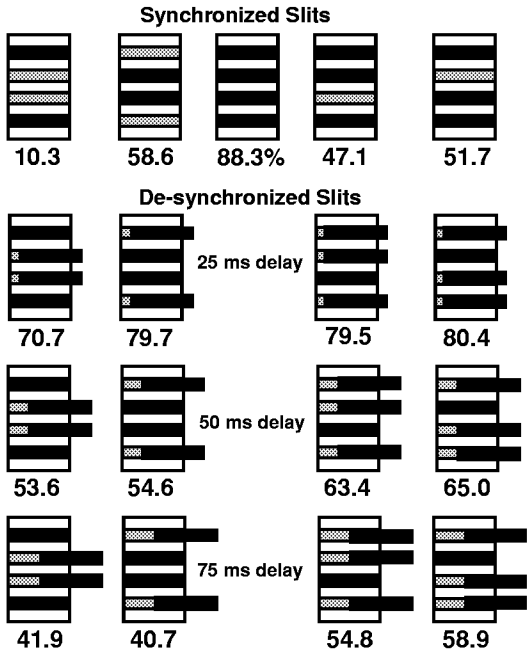

Figure 4. The effect of slit asynchrony on the intelligibility of 4-slit-compound sentences. The intelligibility associated with five baseline conditions is illustrated for comparison. Note that intelligibility diminishes appreciably when the asynchrony exceeds 25 ms, but appears to be *relatively* insensitive to the specific identity of the onset slit(s) (compare left and right adjacent columns).

intelligible). However, intelligibility is unlikely to be based exclusively on this specific parameter of the acoustic signal as it is shared in common with all but the highest-frequency slit when presented alone. Some other parameter (or set) must also be involved.

5. INTELLIGIBILITY DEGRADES WITH SMALL AMOUNTS OF ASYNCHRONY

A second experiment addressed this issue by systematically desynchronizing the 4-slit compound in order to ascertain asynchrony's effect on intelligibility. The results of this experiment (Figure 4) clearly demonstrate the importance of the phase component of the modulation spectrum since even asynchronies as small as 50 ms have an appreciable effect on intelligibility. For this reason, the temporal registration of modulation patterns across spectral frequency channels is likely to play a crucial role in understanding spoken language.

This conclusion is seemingly at odds with the results of a previous study [1] [4] in which 1/4-octave channels of *full-bandwidth* speech were temporally scrambled in quasi-random fashion without serious effect on intelligibility at all but the highest degrees of asynchrony. Fully 80% of the words were correctly decoded even when the channels were desynchronized by 140 ms (average asynchrony = 70 ms, the average length of a phonetic segment in spoken English [5]). In contrast, intelligibility ranged between 40 and 60%

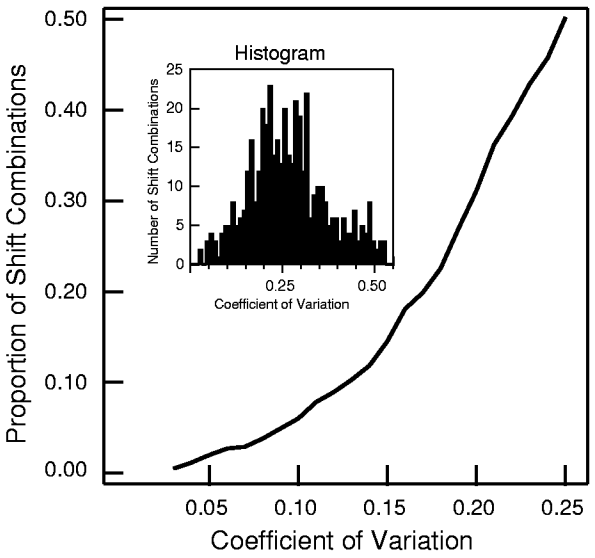

Figure 5. The coefficient of variation (variance/mean, with the offset removed) associated with the range of channel asynchrony characterizing the sentential stimuli used in the experiment described in [1] and [4]. The histogram (insert) illustrates the coefficient of variation's distribution (based on 448 possible channel combinations). The primary plot shows the cumulative distribution for these same data and indicates the presence of ca. 35 channel combinations (8% of the total) distributed across the signal's frequency spectrum with relatively small degrees of asynchrony (c.v. < .1).

for comparable amounts of asynchrony in the current experiment. A potential resolution of this seeming paradox is illustrated in Figure 5. The coefficient of variation (c.v.; the variance/mean, with the offset subtracted) associated with the 448 possible combinations of four 1/4-octave channels distributed over octave sub-regions of the spectrum (the quantizing interval of analysis for this earlier study) spans a very wide dynamic range (0.02 to > 0.5). Small coefficients (< 0.1) reflect very small degrees of asynchrony, on the order of 10 ms or less for an average desynchronization of 70 ms. Approximately 8% of the channel combinations fulfill this criterion. The intelligibility of such channel-desynchronized sentences may therefore be derived from a relatively small proportion of auditory channels strategically distributed across the tonotopically organized spectrum.

6. CONCLUSIONS

A detailed spectro-temporal analysis of the speech signal is not required to understand spoken language. An exceedingly sparse spectral representation is sufficient to accurately identify the overwhelming majority of words in spoken sentences, at least under ideal listening conditions. A more likely basis for spoken language understanding is the amplitude and phase components of the modulation spectrum (cf. [8] for a similar perspective derived from automatic speech recognition studies) distributed across the frequency spectrum. Such a representation would be of utility in improving the technology underlying applications ranging from automatic speech recognition to auditory prostheses for the hearing impaired.

7. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (SBR-9720398) and the International Computer Science Institute. We wish to thank Joy Hollenback and Diane Moffit for assistance in running the experiments and to express our appreciation to the students at the University of California, Berkeley who willingly gave of their time (and ears) to provide the data described. The experimental protocol was approved by the Committee for the Protection of Human Subjects of the University of California, Berkeley.

8. REFERENCES

[1] Arai, T. and Greenberg, S. "Speech intelligibility in the presence of cross-channel spectral asynchrony." *Proc. IEEE ICASSP*, Seattle, pp. 933-936, 1998.

[2] Arai, T., Hermansky, H. Pavel, M. and Avendano, C. "Intelligibility of speech with filtered time trajectories of spectral envelopes." *Int. Conf. Spoken Lang. Proc.*, Philadelphia, pp. 2490-2493, 1996.

[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] Greenberg, S. and Arai, T. "Speech intelligibility is highly tolerant of cross-channel spectral asynchrony." *Proc. Acoust. Soc. Am./Int. Cong. Acoust.*, Seattle, pp. 2677-2678, 1998.

[5] Greenberg, S., Hollenback, J. and Ellis, D. "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus." *Int. Conf. Spoken Lang. Proc.*, Philadelphia, pp. S32-35, 1996.

[6] 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.

[7] 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.

[8] Kanedera, N., Hermansky, H. and Arai, T. "On the properties of modulation spectrum for robust speech recognition," *Proc. IEEE ICASSP*, Seattle, pp. 613-616, 1998.

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

[10] Lippmann, R. "Accurate consonant perception without mid-frequency speech energy." *IEEE Trans. Sp. Aud. Proc.* 4: 66-69, 1996.

[11] 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, pp. 21-52, 1987.

[12] Warren, R. M., Riener, K. R., Bashford, J. A.n and Brubaker, B. S. "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits." *Percept. Psychophys.*, 57: 175-182, 1995.