ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

TEMPORAL CONSTRAINTS ON SPEECH INTELLIGIBILITY AS DEDUCED FROM EXCEEDINGLY SPARSE SPECTRAL REPRESENTATIONS

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

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

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

10.21437/Eurospeech.1999-594

ABSTRACT

A novel means of quantifying the contribution of specific spectral bands for intelligibility is described. The spectrum of spoken English sentences is partitioned into one-third octave bands ("slits") and the contribution of each of four slits ascertained independently and in combination with other slits distributed across the spectrum. The intelligibility baseline (four concurrent slits) yields ca. 85% intelligibility. The current study demonstrates that intelligibility progressively declines as the two central slits (2+3) are desynchronized between 25 and 250 ms. Beyond 250 ms intelligibility often declines even further but then begins to *increase* for greater degrees of asynchrony, suggesting the presence of a perceptual processing buffer of ca. 200-300 ms in duration. The utility of the spectral slit technique is also demonstrated for estimating the contribution towards intelligibility of different regions of the modulation spectrum. The mid-frequency (10-25 Hz) modulations are shown to be of particular significance for encoding speech information above 1.5 kHz. These two experiments demonstrate the power and utility of using circumscribed portions of the spectrum for quantitative evaluation of the contribution made by specific spectro-temporal properties of the speech signal.

1. INTRODUCTION

The high degree of redundancy contained within the speech signal poses a significant challenge for quantifying the contribution of specific spectral components for speech intelligibility. The full-bandwidth spectrum (0.1-6 kHz) can not effectively be used as a performance baseline since its intelligibility is highly robust to various sorts of background interference and distortion. Intelligibility for such a signal is essentially greater than 100% (i.e., there is a considerable "ceiling" effect), thus making it difficult to delineate with precision the contribution made by specific spectro-temporal properties of the acoustic signal towards understanding spoken language.

A novel technique has been developed that provides a method for quantifying the contribution made by specific spectral channels to speech intelligibility.

The spectrum of spoken English sentences (from the TIMIT corpus, incorporating speakers from different dialect regions of the U.S.) was partitioned into 1/3-octave bands ("slits") and the contribution of each of four slits ascertained independently and in combination with other slits distributed across the spectrum [4]. Four slits, distributed between 300 and 6000 Hz typically yield 80-85% intelligibility, providing a performance baseline with which to quantify the contribution of each band towards the total "gestalt" of spoken language comprehension.

2. SIGNAL PROCESSING AND PRESENTATION

Stimuli were sentences read by speakers (of both genders) spanning a wide range of American dialectal regions, age and voice quality. The acoustic signals were initially sampled at 16 kHz, but further low-pass filtered at 6 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 at least an octave in order to minimize intermodulation distortion and masking effects potentially arising from the interaction of non-continuous, spectrally proximal components.

No sentence was presented on more than one trial, minimizing the effects of learning and memorization on intelligibility performance. However, listeners were allowed to listen to any given sentence up to four times before typing in the word sequence. Subjects were paid for their time. All listeners were native speakers of American English with no reported history of hearing loss. The signals were presented by computer and played over high-quality headphones at a comfortable listening level (under subject control) in a sound-attenuated room.

3. DATA COLLECTION AND ANALYSIS

Each listener listened to five practice sentences before beginning the experiment proper. Subjects were instructed to type the sequence of words heard into the computer. For each experimental condition 10 different sentences were presented. Although each subject listened to the same set of sentences as presented to the other subjects, the specific sequence of sentences played was varied so as to minimize the likelihood that variation in intelligibility could be attributed to the identity of specific sentential material.

The number of correct words per sentence was scored using an algorithm that automatically compensated for minor errors in spelling. The proportion of words correctly typed (and in the proper sequence) was computed for each block of ten sentences associated with a specific experi-

mental condition. Subjects who failed to achieve a performance of 70% words correct for the four-slit, synchronous (i.e., no delay) condition were excluded from the data analysis. Virtually all of the listeners correctly transcribed 80% or more of the words in this baseline condition.

4. INTELLIGIBILITY OF ASYNCHRONOUS SLITS

In a previous study [4] it was demonstrated that two central slits (centered around 850 and 2100 Hz) provide by themselves ca. 61% intelligibility, while the lowest (center frequency = 335 Hz) and highest (cf = 5400) slits provide, in concert, only ca. 10% intelligibility (cf. Figure 2). In this earlier study it was also shown that desynchronizing the two central slits (2+3) relative to slits 1 and 4 results in a progressive decline in intelligibility as the asynchrony increases from 25 ms (ca. 72% correct) to 75 ms (ca. 53% correct). In other words, a 75-ms asynchrony results in a level of performance slightly lower than that associated with slits 2 and 3 alone (cf. Figure 2), suggesting some degree of interference between the two pairs of slits.

The current study demonstrates that the intelligibility continues to decline for slit asynchronies up to 250 ms (41% correct), in seeming contradiction to listeners' ability to understand sentential material when full-band speech signals are spectrally desynchronized by up to 140-200 ms [1] [3]. A previous study showed that the apparent tolerance for such spectral asynchrony is the consequence of redundancy in the speech signal which serves to mask the auditory system's sensitivity to the phase of the modulation spectrum across frequency [4]. The precipitous decline of word intelligibility below the baseline for two slits presented in isolation suggests that there is a significant degree of interference between the acoustic image associated with the central channels and those pertaining to the lower and upper spectral regions.

For longer asynchronies (300-600 ms), the intelligibility increases slightly, reaching a mean level of performance of ca. 50% (Figure 1). However, within the averaged intelligibility data is embedded a rather interesting pattern. Virtually all listeners exhibit a non-monotonic function for intelligibility associated with slit asynchronies of 300 ms or longer. The most general pattern manifests a peak of intelligibility at 300 ms (ca. 62% correct) followed by a trough (ca. 43-47% correct) at 400-500-ms asynchrony, with a rise in intelligibility at 600 ms (ca. 58%), close to the baseline performance (Slits 2+3 alone) of 61%. A second pattern exhibits a minimum at 300 ms (ca. 35% correct) with an increase in intelligibility to ca. 50% at 400 and 500 ms (and a slight decline to ca. 40% at 600 ms). Such oscillations in intelligibility are not entirely accountable in terms of the difficulty of the sentential material across conditions. The variation in intelligibility performance appears rather to reflect intrinsic variability in the ease with which sentential material is decoded as a function of spectral asynchrony. The most parsimonious explanation for this pattern of intelligibility is the presence of a perceptual buffer of ca. 200-300 ms that affects a listener's ability to integrate acoustic information across time and frequency.

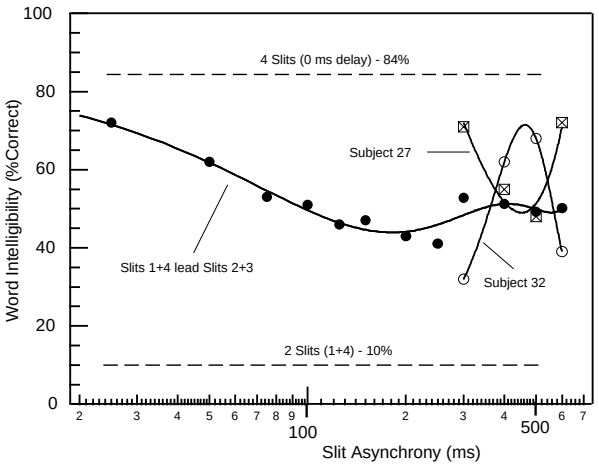

Figure 1 Mean word intelligibility as a function of slit asynchrony (Slits 1+4 leading Slits 2+3). Baseline performance is indicated for several conditions. Two of the 16 subjects' individual data are plotted to illustrate the sorts of representative variation observed for long delays among these listeners.

5. DIFFERENTIAL CONTRIBUTION OF THE MODULATION SPECTRUM ACROSS FREQUENCY

The spectral slit technique can also serve to delineate the intelligibility of different regions of the modulation spectrum across frequency. Rob Drullman and colleagues have suggested that most of the information required for understanding speech is contained in the portion of the modulation spectrum below 8 Hz [2]. In their study the modulation spectrum was low-pass filtered uniformly across frequency. This uniformity of modulation spectral filtering was based on the assumption that the shape of the

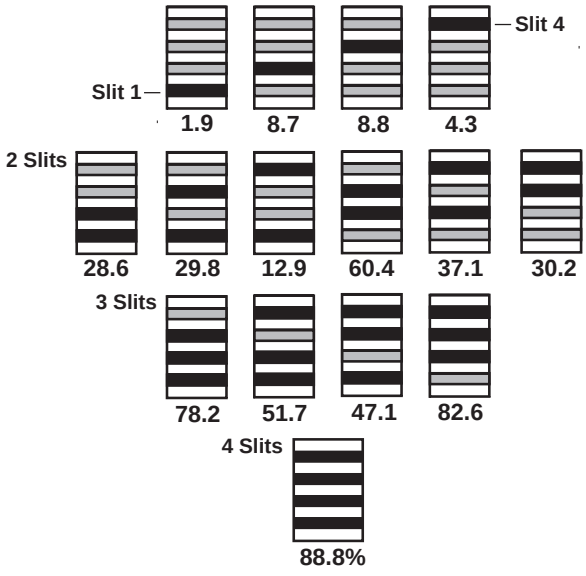

Figure 2 Intelligibility of spectral-slit sentences under 15 separate listening conditions. Baseline word accuracy is 88.8% (4-slit condition). The intelligibility of the multiple-slit signals is far greater than would be predicted on the basis of word accuracy (or error) for individual slits presented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

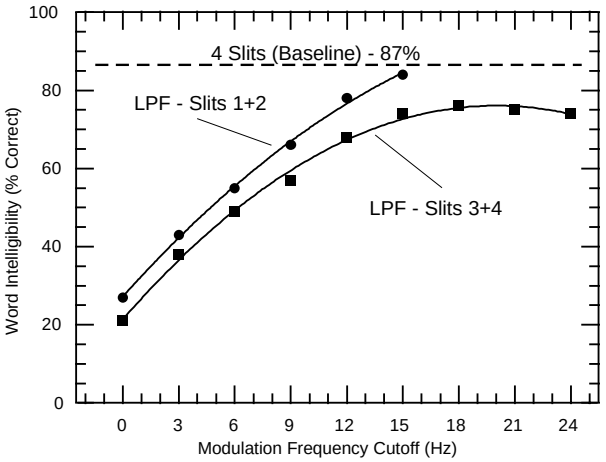

Figure 3 Intelligibility of 4-slit sentential material as a function of the low-pass filter cut-off frequency of the indicated slit(s). Four slits were always presented. In this instance intelligibility of the low-pass-filtered two upper slits is compared with that of the low-pass-filtered lower pair of slits.

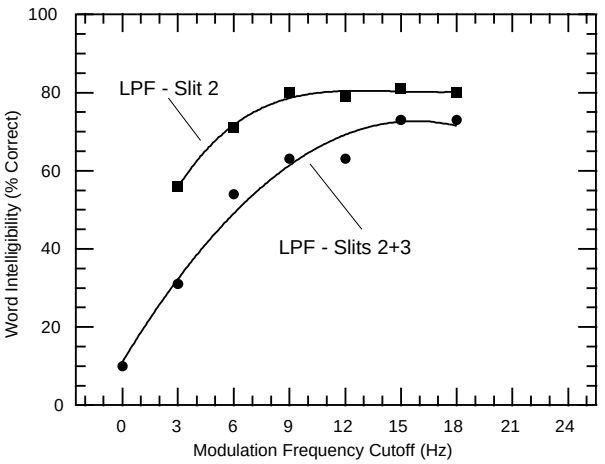

Figure 4 Intelligibility of 4-slit sentential material as a function of the low-pass filter cut-off frequency of the indicated slit(s). The two center slits are low-pass-filtered and the intelligibility compared to sentences in which only Slit 2 is low-pass-filtered. Notice the appreciable performance differential between conditions.hhh

modulation spectrum was relatively similar across frequency [5]. However, it has recently been demonstrated that the higher frequency channels (>1.5 kHz) contain a significantly greater amount of energy in the mid-frequency modulation spectrum (10-25 Hz) than the lower portion of the frequency spectrum, thus raising the possibility that these mid-frequency modulations may play an important role in the coding of certain phonetic and syllabic properties of spoken language.

In the present study all four slits were presented concurrently, but the modulation spectrum of one or two of the slits was low-pass filtered in a systematic fashion. Because the intelligibility for each baseline condition is known in advance (cf. Figure 2) it is possible to directly ascertain the impact of selectively filtering the modulation spectrum of specific slits.

It is of interest to examine in detail the intelligibility patterns delineated in Figures 3 through 6. The two lower slits reach an asymptotic level of intelligibility at ca. 15 Hz (i.e., further increases in the bandwidth of the modulation spectrum does not increase listeners' performance). In contrast, the intelligibility of the two upper slits does not come close to reaching an asymptotic level even when modulation frequencies up to 21 Hz are included (Figure 3). When the modulation pattern of the two central slits are low-pass-filtered intelligibility is still lower than the baseline level (Figure 3). This reduction in intelligibility is likely to be due to Slit 3, in view of the fact that when only Slit 2 is subjected to modulation filtering, listening performance is high (Figure 3).

By comparing intelligibility when Slits 1 and 2 are modulation filtered (Figure 5) with performance when

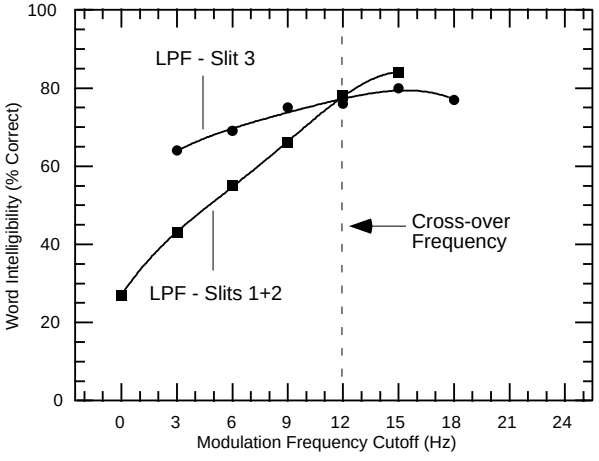

Figure 5 Intelligibility of 4-slit sentential material as a function of the low-pass filter cut-off frequency of the indicated slit(s). The intelligibility of the two lower slits is compared with slit 3. Note that the cross-over point lies 12 Hz, suggesting that the modulation spectrum information is most important below this frequency.

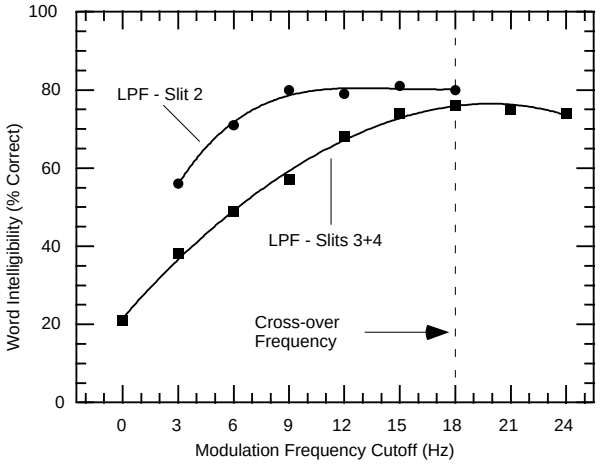

Figure 6 Intelligibility of 4-slit sentential material as a function of the low-pass filter cut-off frequency of the indicated slit(s). The two upper slits are low-pass-filtered and the intelligibility compared to sentences in which only slit 2 is low-pass-filtered. Notice the cross-over point is ca. 18 Hz, in contrast to that of Figure 5.

only Slit 3 is filtered, it is possible to deduce that most of the significant portion of the modulation spectrum for low-frequency channels is below 12 Hz (the "cross-over" point for the two functions in Figure 5).

Analogously, it is possible to infer that the upper frequency channels are more likely to utilize information in the mid-frequency range of the modulation spectrum, given the cross-over point of Figure 6 is located at 18 Hz. These data suggest that there are significant differences in the way in which the modulation spectrum is processed across the spectrum. Such differences in the modulation spectrum could be of importance for assessing the speech reception potential of hearing impaired listeners, who typically experience greatest difficulty with the portion of the spectrum above 1.5 kHz and could also account for the inability of the Speech Transmission Index [5] to successfully predict the speech intelligibility capabilities of hearing-impaired individuals basis on standard psychoacoustic tests.

File	Static Content	Low Pass (Hz)	Changing Content	Delay (ms)
s067.wav	Original			
s067ii.wav	4 Slits			
s0670i.wav	Slits 3+4			
s0671i.wav	Slits 3+4	3	Slits 1+2	
s0673i.wav	Slits 3+4	9	Slits 1+2	
s0676i.wav	Slits 3+4	18	Slits 1+2	
s067i0.wav	Slits 1+2			
s067i1.wav	Slits 1+2	3	Slits 3+4	
s067i3.wav	Slits 1+2	9	Slits 3+4	
s067i6.wav	Slits 1+2	18	Slits 3+4	
s067i0.wav	Slits 2+3			
s067i1.wav	Slits 1+4	3	Slits 2+3	
s067i3.wav	Slits 1+4	9	Slits 2+3	
s067i6.wav	Slits 1+4	18	Slits 2+3	
s067si.wav			Slits 2+3	
s067s1.wav	Slits 1+4		Slits 2+3	25
s067s2.wav	Slits 1+4		Slits 2+3	75
s067s3.wa	Slits 1+4		Slits 2+3	150
s067s4.wav	Slits 1+4		Slits 2+3	300
s067s5.wav	Slits 1+4		Slits 2+3	600

Table 1 Computer waveform files illustrating experimental parameters used in the current study. Refer to text for description of specific parameters of the acoustic signal that are manipulated in the demonstrations. Waveform files are found on the CD-ROM accompanying the 1999 Eurospeech Proceedings (or consult http://www.icsi.berkeley.edu/real/lis). Low past refers to the low-pass cut-off of the modulation spectrum. Delay refers to the asynchrony of the second pair of slits relative to the first. The same sentence is used in all of the demonstrations.

6. CONCLUSIONS

The current study has demonstrated the utility of a novel method with which to ascertain the contribution of specific spectro-temporal components of the speech signal to word intelligibility. By using a calibrated baseline of intelligibility it is possible to deduce the role of specific frequency channels and modulation patterns for understanding spoken language. Through this technique it can be demonstrated that the mid-frequency modulations (10-25 Hz) play a significant role in intelligibility for channels above 1.5 kHz (Figures 3-6). It can also be shown that listeners are exquisitely sensitive to spectral asynchronies as short as 25-50 ms. This phase sensitivity to the modulation spectrum may be of particular importance for the hearing impaired under noisy and reverberant conditions.

7. ACKNOWLEDGEMENTS

The research described in this paper was supported by a grant from the Learning and Intelligent Systems Initiative of the National Science Foundation and the Department of Defense. We thank Joy Hollenback for her assistance in running the experiments, as well as the subjects for their diligence and time. The experimental protocol conformed to the guidelines of the Committee for the Protection of Human Subjects at 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] Drullman, R., Festen, J. M. and Plomp, R. "Effect of temporal envelope smearing on speech reception." *J. Acoust. Soc. Am.*, 95: 1053-1064, 1994.

[3] 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.

[4] Greenberg, S., Arai, T. and Silipo, R. "Speech intelligibility derived from exceedingly sparse spectral information." *International Conference on Spoken Language Processing*, Sydney, 1998, pp. 2803-2806.

[5] 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.