A REVIEW OF SOME PRELIMINARY RESULTS ON THE AUDITORY REPRESENTATION OF SPEECH

M. F. Cheesman, Ph. D.
Department of Communicative Disorders
Elborn College
University of Western Ontario
London, ON
N6G 1R8

ABSTRACT

Speech perception studies have traditionally focused on the identification of the acoustic cues that are needed to distinguish the sounds of speech. Recently, however, techniques have been developed that allow investigation of the auditory speech cues. Direct physiological measurements have been made of the output of the auditory system at the levels of the auditory nerve and cochlear nucleus during stimulation by speech signals (e.g. Sachs & Young, 1979). In the present paper, some studies that have used psychophysical techniques to estimate the output of the human auditory periphery during speech stimulation and comparison with perceptual measures of the same stimuli are reviewed.

RÉSUMÉ

Les études de la perception de la parole se sont jusqu'ici intéressées à l'identification d'indices acoustiques permettant de distinguer les sons de la parole. Cependant, des techniques, récemment mises au point, offrent la possibilité d'évaluer les indices auditifs de la parole. Ainsi, la réponse du système auditif au niveau du nerf auditif et du noyau cochléaire a été mesurée au moyen d'indices physiologique et ce, en présence de signaux de la parole (e.g. Sachs et Young, 1979). Le présent article expose une revue de quelques travaux traitant de mesures physiologiques permettant d'estimer la réponse du système auditif périphérique de sujets humains en présence de stimulations par la parole. La comparaison de ces données aux mesures perceptuelles en présence des mêmes stimuli est également passée en revue.

* The paper is based on an invited talk at the annual meeting of the Canadian Acoustical Association, October 6, 1988.

Traditionally, cues for speech sound recognition have been defined almost exclusively in terms of the acoustic properties of the speech signal itself. The search for the acoustic cues that distinguish the sounds of speech has focused on the systematic manipulation of speech signals in order to determine the effect on perception. This approach has resulted in the generation of a long list of acoustic cues that determine or influence what label a listener assigns to a speech sound. In different listening situations, the relative perceptual weight of each of these cues may change.

For example, some of the salient acoustic cues which have been identified as contributing to the distinction between "gold" and "bold" are the frequency of an initial aperiodic energy burst, the direction of change of the second spectral prominence in the periodic portion of the signal and the overall shape of the spectrum of the first 25 ms or so of the word (e.g. Blumstein & Stevens, 1980; Cooper, Delattre, Liberman, Borst & Gerstman, 1952; Kewley-Port, 1983; Liberman, Delattre, Cooper & Gerstman, 1954). Furthermore, the acoustic information that distinguishes these two speech sounds differs depending on which sounds occur before and after the sound, the talker characteristics of the person producing the sound, and the location of the sound in a word or an utterance (e.g. Delattre, Liberman & Cooper, 1955).

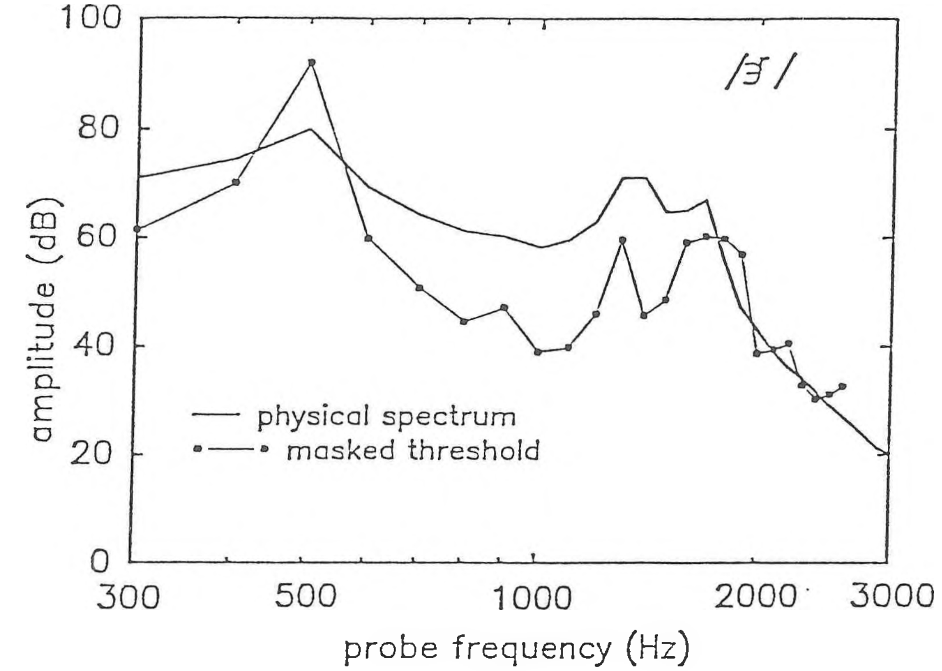
Although such acoustically-based investigations have yielded much information concerning the relevant acoustic parameters of speech, recent contributions from auditory physiology offer the possibility for studying the dependence of speech perception on the auditory representation of speech cues. Of particular interest is the question of what speech spectra would look like at the output of the auditory periphery rather than at the output of a digital-to-analog converter. For example, to what extent do the temporal and frequency resolving powers of the ear preserve the potential speech cues identified by acoustically-based analyses? When a change in the acoustic information in a speech signal does not change the listener's percept, is this because of the lack of a corresponding change in the auditory representation of the signal or because some higher level decision process has determined that both auditory patterns constitute the same speech sound?

The impetus provided by the recent, multi-disciplinary research efforts to understand speech processing has led to two converging approaches to quantifying this auditory form of speech information. The first approach is to formulate explicit models of human auditory processing of speech. These models have been developed by individuals with engineering and/or physiology backgrounds, such as Delgutte (1987), Searle, Jacobsen and Rayment (1979) and Seneff (1987). The models are based on known and predicted physiological responses of the peripheral mammalian auditory system to complex auditory stimuli and, in particular, to speech sounds. The outputs of the mathematical models are then compared to acoustic representations of the signal, such as amplitude spectra.

Empirical studies, such as those undertaken by Carney and Geissler (1986), Delgutte (1980; 1984), and Sachs and Young (1979), have demonstrated that the physical characteristics of the speech signal are well-represented in the auditory-nerve firing patterns, as measured by rate and period synchronization. The auditory nerve functions as a carrier of an internal representation of the temporal and spectral characteristics of the acoustic signal.

Such approaches have provided valuable insights for speech researchers. However, they do not directly measure human processing of speech, and they cannot be used to compare peripheral speech processing and speech perception in the same individual with the same speech token. An alternative research approach is to use psychophysical techniques to estimate the auditory patterns available to the listener when perceiving speech.

The amount of masking produced by a stimulus at different frequencies can provide a psychophysical estimate of the auditory representation of that sound. For example, using a steady-state vowel as a masker, a pattern of the masked thresholds for each of a series of probe tone frequencies can be obtained. In figure 1, the spectrum of an /ɝ/ vowel, used as a masker by Van Tasell, Fabry and Thibodeau (1987), is shown by the solid line. Three spectral prominences are present in this steady-state, synthetic vowel, corresponding roughly to the first three formants of the vowel masker. Along with this physical spectrum, transformed forward-masked thresholds for probes ranging from 300 to 2600 Hz are plotted with connected filled circles. The ordinate represents the amount of masking produced by the vowel.



Adapted from Van Tasell, Fabry and Thibodeau (1987).

Figure 1. Solid line: spectrum of a steady-state synthetic /ɝ/ vowel that was used as the masker. Data points (filled circles) represent transformed, forward-masked thresholds.

This masking pattern is taken to be a frequency-domain representation of the auditory excitation pattern produced by the vowel. Not surprisingly, vowel masking patterns such as these have shown that the auditory representation of a vowel may differ from its acoustic spectrum in ways that are consistent with what is known about the processing capabilities of the peripheral auditory system (Bacon & Brandt, 1982; Moore & Glasberg, 1983; Van Tasell, Fabry & Thibodeau, 1987). In particular, the spectral peaks in the masking pattern are sharper than in the acoustic spectrum, due to the suppression of masker activity in frequency regions that are adjacent to the spectral peaks. These suppression effects are only visible in non-simultaneous masking conditions, in which the masker activity may suppress itself, but not the probe activity.

In the case of a dynamic speech signal, such as a consonant-vowel syllable, the time frame over which the output of the auditory periphery is observed can be manipulated by changing the duration of a probe tone and assuming that the masked threshold represents the pattern of activity integrated over the duration of the probe tone. Unlike traditional acoustical analyses, the effect of temporally adjacent acoustic information on the auditory representation will be included in the estimate of the auditory representation. The nonsimultaneous masking effects of adjacent speech segments will influence the detection of the probe tone and will therefore be reflected in the masking patterns. This is a further advantage of the auditory over acoustic modes of representing complex sounds.

Two studies in which this psychophysical masking paradigm was used to estimate the auditory representation of speech in human listeners differ from most similar studies because they also included perceptual measures obtained with the same speech signals. Cheesman, Van Tasell and Ortmann (1987) reported the results of a study in which the speech maskers were natural consonant-vowel syllables, /ba/, /da/ and /ga/, spoken by a female talker. The first 25.6 ms of these signals, representing the portion of the consonant that contributes the most to the identification of the consonant, was subjected to linear predictive coding (Markel & Gray, 1976). The results of these analyses for the three maskers are shown by the solid lines in Figure 2.

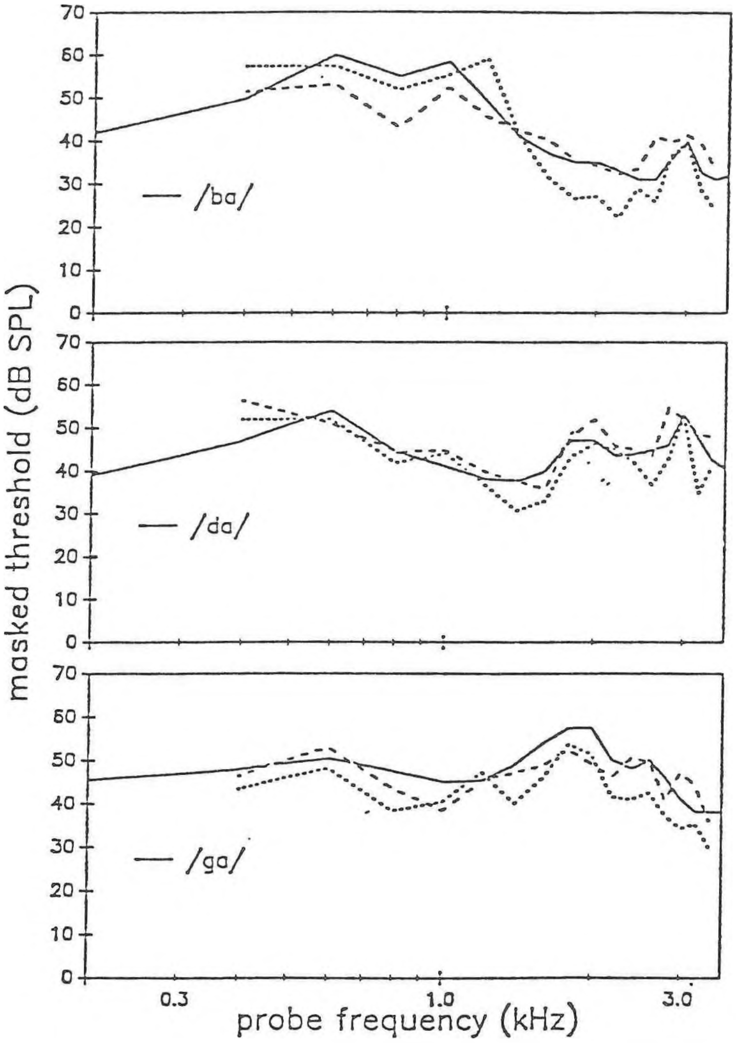


Figure 2. Solid lines represent the LPC spectra of the first 25.6 ms of the masker syllables /ba/, /da/, and /ga/, in the top, middle and bottom panels,. respectively. Dashed lines (S1) and dotted lines .(S2) connect the masked thresholds for 20-ms probes presented at the onset of the masker syllables.

Psychophysical masking patterns were obtained for probe tones that were presented during the first 20 ms of the masker syllables. Thresholds were estimated with a two-interval forced-choice adaptive-tracking procedure (Levitt, 1971) and were an average of thresholds obtained over two adaptive runs.

The masked thresholds for the probes for the two listeners are represented by the broken lines in Figure 2. Not only did these masking patterns match the overall shape of the onset spectra, but also most of the fine structure remained as well. For the /ba/ and /da/ maskers -- in the top and middle panels -- masked thresholds were highest at or near the frequencies corresponding to the first three spectral peaks. Van Tasell, Fabry and Thibodeau (1987), using steady-state synthetic vowels as maskers in a forward masking paradigm, have also reported that the frequency of masking pattern peaks may be shifted with respect to the peaks in the acoustic spectrum of the masker. This result is shown in the data of Cheesman, Van Tasell and Ortmann (1987) for listener S2, for whom the second peak in /ba/ was shifted up in frequency in the masking pattern relative to its position in the physical spectrum. With the /da/ masker, the third spectral peak was shifted down for the listener Sl.

For the masker /ga/, in which the second and third spectral peaks of the masker were relatively close in frequency, the peaks in the masking pattern were less systematically related to the masker spectrum. The two listener's patterns differed from each other and from that of the onset spectrum in both the number and location of the maxima.

Cheesman (1989) also obtained masking patterns using dynamic speech maskers. However, in this study masking patterns were obtained at several points in time during the speech masker. The masking stimuli were synthetic syllables that were perceived by the subject as being the consonants "s" or "sh" paired with the vowels "i" and "u". The consonants "s" and "sh" are typically characterized by several hundred milliseconds of bandpass noise that is generally higher in frequency for "s" sounds than for "sh" sounds.

The stimuli chosen for this study are particularly interesting because the same aperiodic energy is perceived as a different consonant, either "s" or "sh", depending on which of the two vowels follow it. Figures 3 and 4 contain spectrographic displays (CSRE, Jamieson, Nearey, & Ramji, 1989) of the two masking syllables. The noise spectra were identical for the two syllables; because the syllables were formed with the same frozen noise, only the vowels differed. Figure 3 depicts the noise followed by an /i/ vowel of the syllable that was identified as the word "she" by the listener; Figure 4 depicts the same noise followed by an /u/ vowel for the syllable that was identified as the word "sue" by the listener.

For each masker, auditory masking patterns were obtained at four points in the syllable: at the intersection of the consonant with the vowel (designated time 0), 25 ms into the vowel (+25 ms), and 25 and 50 ms before the onset of the vowel (-25 and -50 ms, respectively). At each of these points, simultaneous masked thresholds were determined for 10-ms probes at frequencies from 1000-5000 Hz, in 250-Hz steps. Thresholds in this study were made by method of adjustment; that is, the listener adjusted the level of the probe until it was just audible (cf. Spiegel, 1987).

Figures 5 and 6 illustrate the masking patterns produced by two maskers at the four different positions in the syllables. These are data from a single subject. The masked thresholds have been connected by the solid lines. Each data point represents the mean of five threshold adjustments; the standard deviation of the estimates is less than 2 dB. Quiet thresholds for these 10-ms probes,

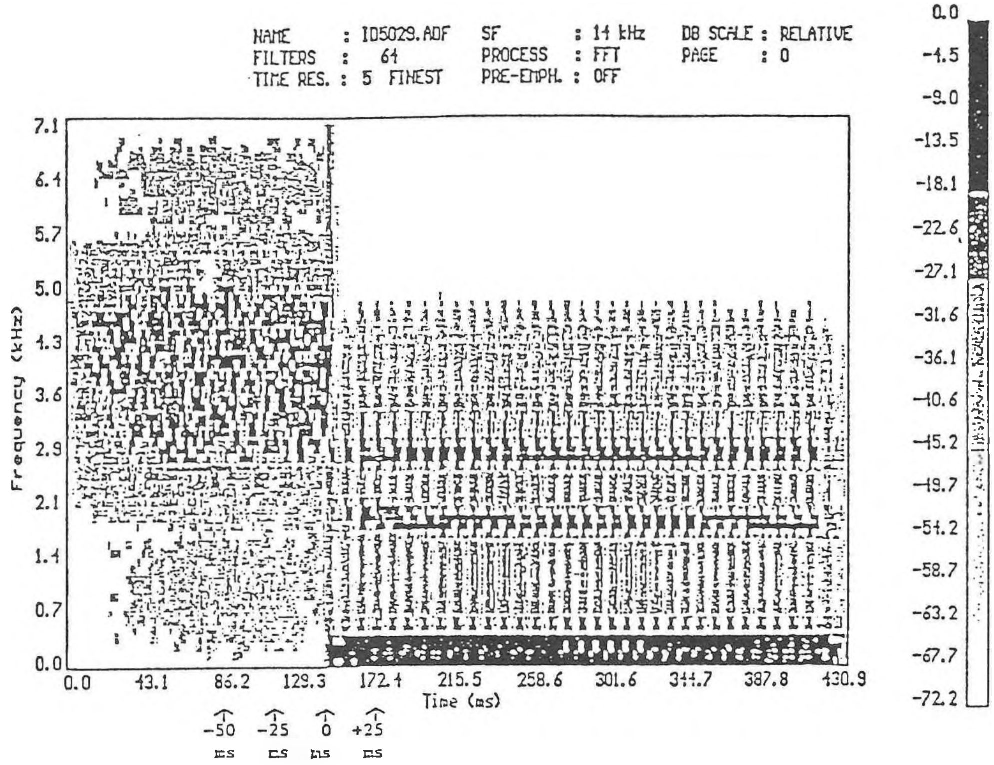


Figure 3. Spectrogram of the synthetic masker syllable "she". Arrows below the spectrogram indicate the four probe positions used to obtain masking patterns.

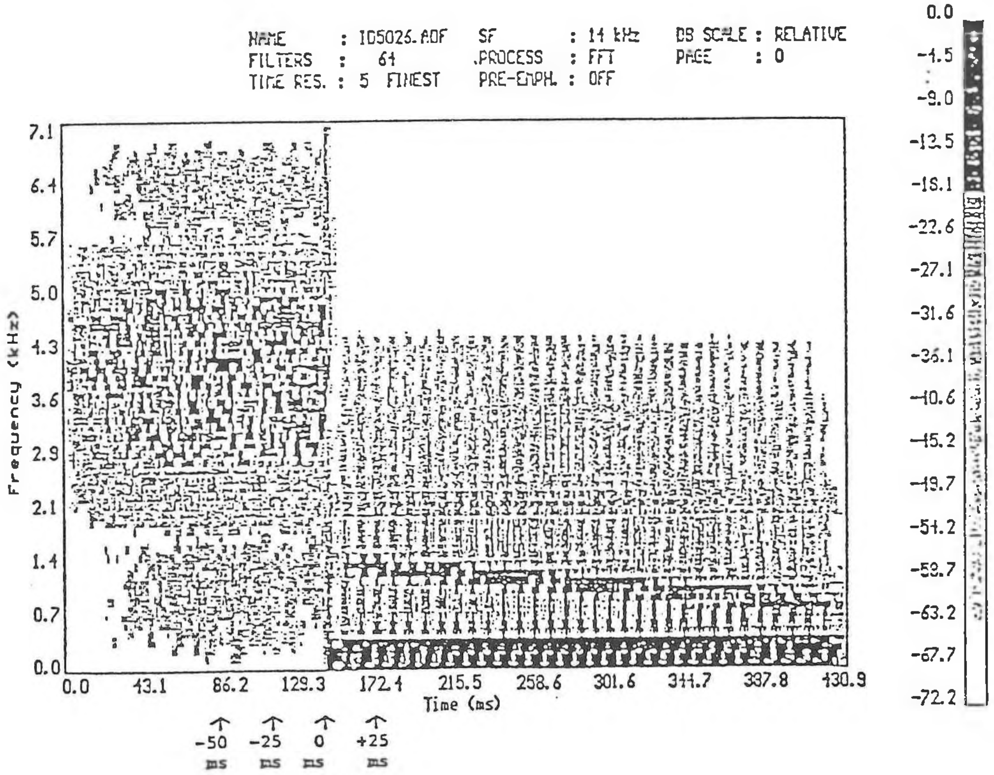


Figure 4. Spectrogram of the synthetic masker syllable "sue".

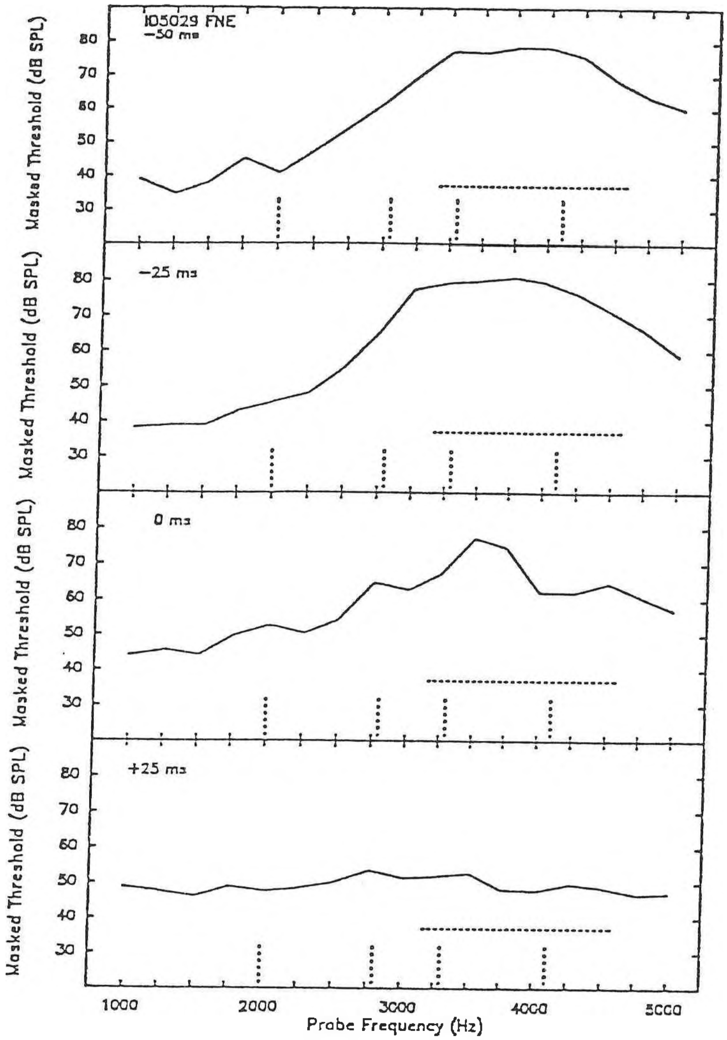


Figure 5. Masked thresholds as a function of probe frequency for 10-ms probes at each of four positions in the "she" masker. The dotted horizontal bar represents the passband of the consonant noise. The dotted vertical lines indicate the frequencies of the spectral prominences in the vowel between 1 and 5 kHz.

using the same method of adjustment, were typically around 28 db SPL for frequencies below about 3 kHz and rose to around 35 dB SPL at the highest frequencies used in this study.

In addition to the masked thresholds, some parameters of the masking stimuli have been plotted on these curves. The dotted horizontal line indicates the passband of the consonant noise. The four vertical lines at the bottom of each panel indicate the frequencies of the spectral prominences in the vowel portion of the maskers. Recall that during the -25 and -50 ms probe times, only consonant noise is present in the masker. At time 0, the consonant ceases and the vowel commences and the energy present is quite low. At +25 ms, only vowel energy is present in the masker.

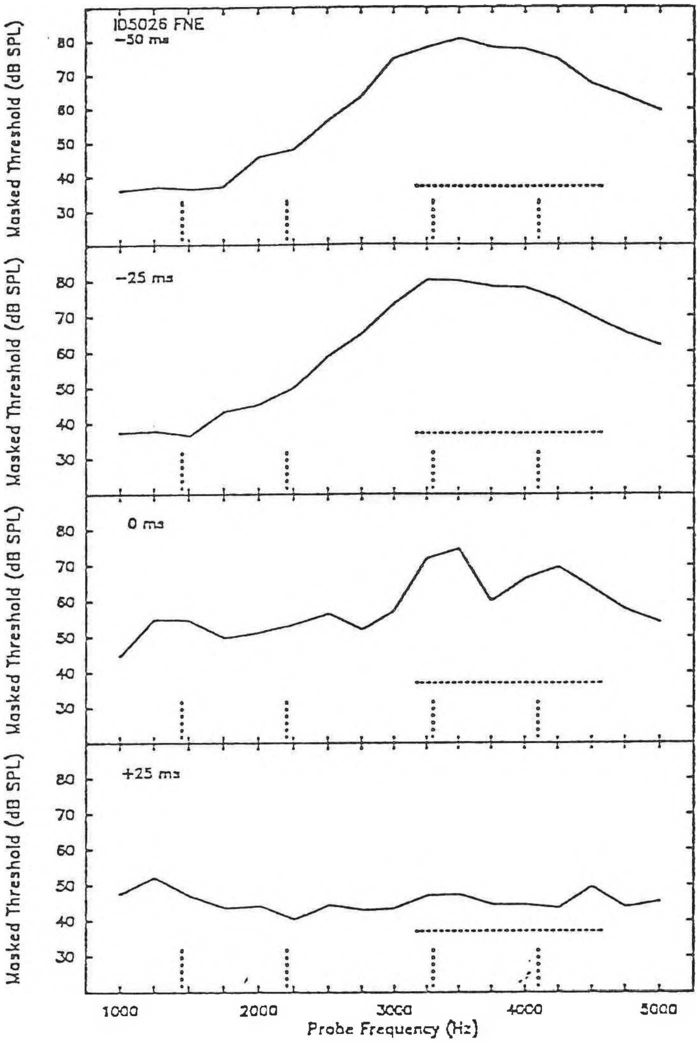


Figure 6. Masked thresholds as a function of probe frequency for 10-ms probes at each of the four positions in the "sue" masker. Dotted lines represent masker parameters, as in Figure 5.

Three points are noteworthy in these data. First, during the consonantal portion of the masker, most of the masking occurred at the frequency of the consonant. Second, there was some indication during the consonant and at the intersection of the consonant with the vowel (time 0) that the vowel itself was doing some backward masking, i.e. raising the threshold in frequency regions where there were peaks in the vowel spectrum.

Third, there was a relatively flat masking pattern during the initial part of both of the vowels (+25 ms), in which some small peaks were observable that probably correspond to peaks in the vowel. More importantly, these flat patterns provided an example of a limitation of the simultaneous masking technique in the estimation of the auditory representation of dynamic signals. Although there were more intense spectral peaks in the masker at +25 ms than at 0 ms, this was not reflected in the masking patterns. Masker activity may have effectively suppressed both the masker and the probe and reduced the size of the observed peaks at +25 ms. At 0 ms and earlier, much of masking at the formant frequencies may be backward masking. Such auditory interactions of temporally-adjacent speech energy may be responsible for the perceptual effects described above, in which the perceptual identity of the consonant (either "s" or "sh") was determined jointly by the noise spectrum and by the spectrum of the following vowel.

These research efforts have focused on the relationship between the perceptual identity of a few, dynamic speech sounds, and the auditory representation of those same sounds, as estimated from psychophysical masking patterns. The present efforts are unique in this focus on the relation between these aspects. In future work, this program of research will explore the limitations and interpretations of the estimation procedure and further define the association between the auditory representation of speech and its perception. It is hoped that this work, together with attempts to model auditory peripheral function based on research in auditory physiology, will converge to provide an improved understanding of auditory functioning with speech and other complex acoustic signals.

REFERENCES

Bacon, S. P. & Brandt, J. F. (1982). Auditory processing of vowels by normal-hearing and hearing-impaired listeners. **Journal of Speech and Hearing Research, 25**, 339-347.

Blumstein, S. & Stevens, K. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. **Journal of the Acoustical Society of America, 67**, 648-662.

Carney, L.H. & Geissler, C. G. (1986). A temporal analysis of auditory-nerve fiber responses to spoken stop consonant-vowel syllables. **Journal of the Acoustical Society of America, 79,** 1896-1914.

Cheesman, M. F. (1989). The auditory representation of context-conditioned fricatives. Unpublished doctoral dissertation, University of Minnesota.

Cheesman, M. F., Van Tasell, D. J. & Ortmann, T. M. (1987). Masking patterns for syllable-initial stop consonants. ASHA Annual Meeting, New Orleans.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M. & Gerstman, J. L. (1952). Some experiments on the perception of synthetic speech sounds. **Journal of the Acoustical Society of America, 24**, 597-606.

Delattre, P. C., Liberman, A. M. & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. **Journal of the Acoustical Society of America, 27**, 769-773.

Delgutte, B. (1980). Representation of speechlike sounds in the discharge patterns of auditory-nerve fibers. **Journal of the Acoustical Society of America, 68**, 843-857.

Delgutte, B. (1984). Speech coding in the auditory nerve: II. Processing schemes for vowel-like sounds. **Journal of the Acoustical Society of America, 75**, 879-886.

Delgutte, B. (1987). Peripheral auditory processing of speech information: implications from a physiological study of intensity discrimination. In M. E. H. Schouten (Ed.), **The Psychophysics of Speech Perception**. Dordrecht: Martinus Nijhoff.

Jamieson, D. G., Nearey, T. M. & Ramji, K. (1989). CSRE: A Speech Research Environment. **Canadian Acoustics 17** 23-35.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. **Journal of the Acoustical Society of America, 73**, 322-335.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. **Journal of the Acoustical Society of America, 49**, 467-477.

Liberman, A. M., Delattre, P. C., Cooper, F. S. & Gerstman, J. L. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. **Psychology Monographs, 68**, 1-13.

Markel, J. D. & Gray, A. (1976). **Linear Prediction of Speech**. New York: Springer-Verlag.

Moore, B. C. J. & Glasberg, B. R. (1983). Masking patterns for synthetic vowels in simultaneous and forward masking. **Journal of the Acoustical Society of America, 73**, 906-917.

Sachs, M. B. & Young, E. D. (1979). Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate. **Journal of the Acoustical Society of America, 66**, 470-479.

Searle, C. L., Jacobsen, J. & Rayment, S. G. (1979). Stop consonant discrimination based on human audition. **Journal of the Acoustical Society of America, 65**, 799-809. Seneff, S. (1987). A model for the transduction stage of auditory speech processing. **Journal of the Acoustical Society of America, 82** (Sl), 583(A).

Spiegel, M. F. (1987). Speech masking I. Simultaneous and nonsimultaneous-masking within stop /d/ and flap /r/ closures. **Journal of the Acoustical Society of America, 82**, 1492-1502.

Van Tasell, D. J., Fabry, D. A. & Thibodeau, L. M. (1987). Vowel masking patterns and vowel recognition by hearing-impaired subjects. **Journal of the Acoustical Society of America, 81**, 1586-1597.

ACKNOWLEDGEMENTS

This work was conducted in collaboration with D. J. Van Tasell in the Department of Communication Disorders at the University of Minnesota. Financial support for this research was provided by an SSHRC doctoral fellowship and the Bryng Bryngelson Research Fund.

LARSON-DAVIS LABORATORIES

We have become a new technology leader in acoustics and vibration measuring instruments. Our goal is to provide advanced, precise, high-quality products at very reasonable prices. As the result of a substantial ongoing research program, Larson-Davis products provide versatility and automation untouched by *any* competitive offerings. Our growing product family includes:

- Portable multichannel Real-Time analyzers delivering 1/1, 1/3 and 1/12 octave bands to 125 KHz with future plug-in modules for FFT, acoustic intensity, memory expansion, etc.
- Underwater acoustic analysis equipment.
- Precision sound level meters with computer interfaces and automated control of 1/1 and 1/3 octave filters.
- Data logging noise dosimeters and hand-held sound level meters.
- Environmental and industrial noise monitoring systems.
- Building and architectural acoustics analyzers.
- Vibration measuring and monitoring instruments.
- Audiometric calibration instruments for speech and hearing.
- Network airport noise monitoring systems, with management planning software.
- Precision measuring microphones, preamplifiers, power supplies, instrumentation amplifiers, acoustic intensity probes, calibrators and accessories.

For more information contact the factory.

LARSON • DAVIS
LABORATORIES
280 South Main
Pleasant Grove, UT 84062
801-785-6352 TELEX 705560

Dalimar Instruments Inc.
C.P. 110
Ste-Anne de Bellevue
(Québec) Canada H9X 3L4
(514) 453-0033
TELEX: 05-560592 - TO: 23218

Yvon J. B. Larose
Directeur général
FAX:
(514)453-0554