FLUCTUATIONS IN AMPLITUDE AND FREQUENCY ENABLE INTERAURAL DELAYS TO FOSTER THE IDENTIFICATION OF SPEECH-LIKE STIMULI

Richard M. Stern[1]
Constantine Trahiotis[2]
Angelo M. Ripepi[1]

[1]Department of Electrical and Computer Engineering
and Biomedical Engineering Program
Carnegie Mellon University, Pittsburgh, PA 15213
[2]Department of Neuroscience
and Department of Surgery (Otolaryngology)
University of Connecticut Health Center
Farmington, Connecticut 06030

July 7, 2004

Richard Stern can be reached at (412) 268-2535
Internet: `rms@cs.cmu.edu`

FLUCTUATIONS IN AMPLITUDE AND FREQUENCY ENABLE INTERAURAL DELAYS TO FOSTER THE IDENTIFICATION OF SPEECH-LIKE STIMULI

Richard M. Stern, Constantine Trahiotis, and Angelo Ripepi

ABSTRACT

In this study, we describe the results of two experiments that help clarify the conditions under which interaural time delays can facilitate the identification of simultaneously-presented vowel sounds. In one experiment we measured the intelligibility of simultaneously-presented natural speech and speech that had been degraded in a manner that precluded the use of pitch information. In a second experiment we measured the identification accuracy gained by adding pitch and amplitude information to whispered vowel-like sounds. The major results of these experiments are twofold. First, interaural time delays can indeed facilitate the identification of simultaneously-presented speech-like sounds, even when cues based on common fundamental frequency are not available. Second, the ease with which the very potent contribution of interaural timing information can be exploited is strongly facilitated in turn by the presence of dynamic variations in the stimuli (such as the monaural amplitude and frequency fluctuations that are characteristic of natural speech sounds).

Introduction

This presentation and paper are concerned with the ways in which the use of cues based on interaural time delay (ITD) can foster the separation and identification of multiple simultaneously-presented speech-like stimuli. We begin with a review of some of the major trends of research in binaural modeling, particularly with regard to models based on the interaural cross-correlation of the auditory-nerve response to the stimuli. We then present and discuss the results of two experiments that serve to clarify the extent to which differences in ITD can facilitate the separate perception of competing streams of natural speech, and the important role that is played by the presence of dynamic variations in these stimuli.

The “modern era” of binaural modeling can be said to have begun with Jeffress’s prescient paper (Jeffress, 1948) suggesting that a neural coincidence mechanism could underlie human ability to process small interaural time differences in impinging stimuli. Jeffress suggested that central units could internally code external interaural delays that record coincidences of neural impulses from pairs of more peripheral nerve fibers tuned monaural to the same center frequency. Since then, numerous behavioral and physiological investigations have confirmed the utility and general validity of his observation. Current models include Jeffress’s coincidence mechanism after peripheral processing of the signals that includes bandpass filtering, nonlinear rectification, and compression.

An intuitive understanding of how such cross-correlation-based models can be used to understand the binaural representation of signals presented with external delays can be obtained by considering a three-dimensional “cross-correlation surface.” Figure 1 is a representation of the relative response of an array of coincidence counting units plotted when the signal is a bandpass noise having a center frequency of 500 Hz, a bandwidth of 800 Hz, and presented with an external interaural delay of –1.5 ms. The responses are plotted as a simultaneous function of internal interaural delay (along the horizontal axis) and center frequency of the peripheral input fibers (along the oblique axis). As can be seen, the value of the external delay, –1.5 ms, can be inferred from the “straight” or vertical ridge of the pattern of cross-correlation observed at that internal delay. The remaining, curved, ridges are observed because the cross-correlation functions of the narrowband outputs resulting from peripheral filtering have repetitive peaks of activity that are spaced at intervals of internal delay equal to the reciprocal of the center frequency of the peripheral bandpass filters. Similarly, speech or speech-like information stemming from an external source of sound would be expected to result in a straight ridge of activity corresponding to the internal delay associated with the external delay with which the signal arrives at the two ears. Within this framework the ability to "separate" or understand two sources of auditory information presented simultaneously can be thought about as the ability to parse and/or track the information that corresponds to the respective ridges in the cross-correlation produced by the sources.

Consistent with this view, it is widely accepted that the ability to understand speech in the presence of competing sounds improves when the speech and competing sounds are spatially separated. Such an outcome has been observed in experiments using natural free-field stimuli (*e.g.* Bronkhorst and Plomp, 1990; Yost *et al.,* 1996) and in experiments utilizing stimuli presented via earphones (*e.g.* Bronkhorst and Plomp, 1992; Nilsson and Soli, 1994; Koehnke and Besing, 1996; Yost *et al.,* 1996; Peissig and Kollmeier, 1997; Hawley *et al.*, 1999).

These observations notwithstanding, the results of several experiments using earphones indicate that listeners are unable to acheve separate identification of simultaneously-presented vowel-like sounds solely on the basis of their ITDs (*e.g.* Culling and Summerfield, 1995; Hukin and Darwin, 1995; Darwin and Hukin, 1997). Nevertheless, and important for our purposes, Darwin and Hukin did note that the presence of an ITD can enhance the identification of vowel-like sounds when there is plausible independent evidence of an additional sound source. Taken at face value, the inability to identify artificially-generated vowels on the basis of ITD appears to be inconsistent with the ease with which one can simultaneously perceive naturally-occurring speech emanating from spatially-separated sources.

This investigation is an attempt to understand the factor(s) underlying the differences between these two types of outcomes. Toward that end we conducted two separate, but related, experiments. In the first experiment we attempted to reduce or eliminate pitch-based cues which are inherent in natural speech sounds. This tested the hypothesis that pitch cues could be highly salient, if not necessary, for the identification for the identification of sources of sounds having differing ITDs. In the second experiment we attempted to assess the extent to which the addition of amplitude and/or frequency modulation could combine with ITD to enhance the identification of bandpass-filtered vowel sounds, using stimuli similar to those employed by Culling and Summerfield. It is well known that (monaural) pitch and amplitude cues foster the perceptual segregation of independent sources of sound (*e.g.* Bregman, 1990; Yost, 1992; Darwin and Carlyon, 1995), and it seemed reasonable to expect that the separate perception of simultaneously-presented sources of sounds based on ITD could be facilitated by the presence of monaural modulations in amplitude and/or frequency.

The results of Experiment 1 indicate that the identification of simultaneously-presented natural speech sounds is improved when the ITDs with which they are presented are different, even when pitch information is not available as a cue. The results of Experiment 2 indicate that the addition of speech-like variations in amplitude and frequency can improve the ability to use ITDs in order to identify vowel-like sounds similar to those employed by Culling and Summerfield. This latter finding is consistent with some of the observations of Darwin and Hukin (1997).

Experiment 1: Intelligibility of interaurally-delayed natural speech sounds devoid of pitch information

In this experiment we measured the extent to which pitch information plays a role in the perceptual segregation of interaurally-delayed natural speech. Said differently, we measured intelligibility of speech after removing potentially useful information concerning pitch and harmonicity.

A. Stimuli and experimental procedure

The stimuli were taken from the SATASK database recorded by the US Army Research Laboratory (Koehnke and Besing, 1996). This database consisted of recorded sentences of speech of the form “(NAME) write the number (NUMBER) on the (COLOR) (OBJECT)”. Four names, eight numbers, eight colors, and nine object names were used, all of which were monosyllabic. The sentences were spoken by four males and digitally recorded with a sampling rate of 11,025 Hz under carefully controlled conditions. Efforts were made to ensure that all sentences were spoken at the same rate so that the major content words would occur at coincident times if the signals were combined.

Pairs of sentences selected randomly from the possible combinations of four names, eight numbers, eight colors, and nine object names were combined digitally and presented binaurally. One of the two sentences always began with the NAME “Troy” and will be referred to as the target sentence. The other sentence will be referred to as the masker sentence. An example of a target sentence could be “Troy, write the number 4 on the green fork”. A corresponding masker sentence could be “Ron, write the number 2 on the black kite”. The two sentences were combined with a target-to-masker ratio of 0 dB. In some blocks of trials both sentences were presented with zero ITD, causing both sentences to be perceived in the center of the head when the stimuli were presented diotically over headphones. In other blocks of trials one of the two sentences would be presented with zero ITD while the other would be presented with a 363-μs ITD, causing the dichotically-presented sentence to be perceived toward the leading (right) ear. (With a sampling rate of 11,025 Hz, 363 ms is the closet integer sample delay to the ITD of 400 ms that had been used in the experiments of Culling and Summerfield.)

The pairs of sentences were presented in blocks of 25 trials. They were always composed of speech from two different talkers, and the particular combination of NAME, NUMBER, COLOR, and OBJECT used within each target-masker set was unique. Using a computer terminal with a graphical user interface, the listener's task was to choose the NUMBER, COLOR, and OBJECT that corresponded to the sentence that started with the name TROY. The listeners were required to respond on each trial before the next trial could begin. One hundred sentences from each condition were presented to each of two listeners.

The sentences were presented in five different fashions: natural, vocoded with natural pitch contours, vocoded in monotone style with the target and masker presented at two fixed fundamental frequencies (90 and 100 Hz, respectively), vocoded in monotone style with both target and masker presented with the same fundamental frequency (100 Hz), and whispered. The term "natural" refers to the original speech as recorded in the SATASK database. The various "vocoded" and "whispered" conditions were obtained using LPC waveform coding methods. The incoming speech was windowed using a series of 20-ms Hamming windows which overlapped by 10 ms. Fourteen LPC coefficients were obtained for each windowed segment using the Levinson-Durbin method (*e.g.* Rabiner and Schafer, 1978). These coefficients characterize the time-varying spectral profile of the incoming speech, but do not contain detailed information about the excitation signal. The monotone speech signals were obtained by exciting the linear filter specified by the (time-varying) LPC coefficients with a periodic impulse train. The vocoded speech with natural pitch contours was also obtained by exciting the LPC-derived filter with an impulse train, but with an instantaneous frequency that equaled the fundamental frequency of the voiced segments of the original signal. Pitch was estimated using the pitch-extraction algorithm of the commercially-available Entropic Signal Processing System (Talkin, 1995). The "whispered" speech was obtained by exciting the filter specified by the LPC coefficients with white noise. The use of the monotone and whispered speech conditions enabled us to present speech-like stimuli for which grouping cues associated with fundamental frequency were either difficult to separate (as in the case of two sentences of monotone speech presented simultaneously with precisely the same fundamental frequency), or non-existent (as in the case of simultaneously-presented whispered speech).

B. Results and discussion

Figure 2 displays the percentage of words correctly identified for each of the five experimental conditions, presented separately for two listeners, the first and third authors. The reader is reminded that the masker sentence was always presented with zero ITD. The darker and lighter bars represent data obtained when the target ITD was 0 ms or 363 ms, respectively. Because there were 8 numbers, 8 colors, and 9 objects to identify, chance performance for this task is about 12 percent correct.

For our purposes the most important feature of the data is that identification accuracy was consistently better for target stimuli presented with the ITD of 363 ms. This verifies that ITDs per se can, under appropriate circumstances, facilitate the intelligibility of competing speech-like stimuli. In addition, performance did not vary greatly over the five different types of stimuli. Note that, in particular, a binaural advantage was observed even for conditions in which the fundamental frequencies were the same (the “Same F0” condition) or nonexistent (the “Whispered” condition). We interpret this outcome to mean that ITDs can aid speech intelligibility even when the sources of speech cannot be segregated on the basis of information stemming from variations of pitch. It is interesting listeners were able to perform much better than chance (12 percent) when both targets and maskers were presented with zero ITD. We believe that this is a consequence of the limited masking effects that result when the targets and maskers are presented at an energy ratio of 0 dB.

II. Experiment 2: Identification of simultaneously-presented “whispered” vowels having speech-like contours of pitch and amplitude

A. Stimuli and experimental procedure

The purpose of this experiment was to evaluate the extent to which modulations in amplitude and/or frequency can facilitate the identification of speech-like sounds based on ITD. As in the previous experiment, we employed stimuli that were “intermediate” between natural speech and the whispered vowels used by Culling and Summerfield. Data were also obtained with such modulations absent.

The stimuli included whispered vowels similar to those generated by Culling and Summerfield (1995) as well as versions of such signals that were modulated by the amplitude and/or frequency contours of natural speech. Each of four whispered vowels (which were labeled “AR”, “EE”, “OO”, and “ER”) were constructed by passing white noise through time-invariant filters each of which had two narrow rectangular passbands. Four finite-impulse-response equiripple filters were employed that were obtained using the Parks-McClellan algorithm (*e.g*. Oppenheim and Schafer, 1999). The center frequencies of the passbands were the four center frequencies used by Culling and Summerfield: 225, 625, 975, and 1925 Hz. The filters had transitional bandwidths of 50 Hz and a length of 512 samples. The sample rate was 11,025 Hz, as in Experiment 1. For each trial, independent tokens of each type of stimulus were generated by exciting the appropriate filter with a statistically-independent white noise excitation function.

Contours of amplitude modulation of speech were obtained by measuring the short-term energy of speech waveforms from the Koehnke and Besing SATASK database. Sentences were selected from that database in the same manner as described in the previous experiment. Contours of frequency modulation were obtained by estimating the pitch of the SATASK sentences, again using the pitch-extraction algorithm in the ESPS package of Entropic Research Laboratory. The duration of the sentences was approximately three seconds.

Data were collected using unmodulated whispered vowels, whispered vowels with natural amplitude-modulation contours, whispered vowels with natural frequency-modulation contours, and whispered vowels presented with both natural amplitude and frequency modulation contours. Gaussian noise presented at –30-dB re the level of the vowel sounds was added in order to mask low-level off-frequency sideband information resulting from the modulation of the signals. The unmodulated whispered vowel sounds were similar to the stimuli used by Culling and Summerfield, but approximately three seconds in duration.

For each block of trials, two equal-level vowel waveforms, either an "AR" and an "EE" or an "OO" and an "ER" were combined digitally. One of the two vowels was presented with zero ITD and would be perceived, in isolation, in the center of the head. The other vowel was presented with a 363-μs ITD, and would be perceived, in isolation, toward the right ear. The task was to identify which of the four types of vowels, "AR", "EE", "ER", or "OO", was perceptually toward the right side of midline. As authors and experimenters, the listeners knew that only four combinations of vowel identity and intracranial position were possible, and this information could sometimes be used to improve identification accuracy. This limitation of stimulus conditions was necessary to avoid the presentation of pairs of vowels having overlapping spectral components.

Data were collected in four blocks of 25 trials for each stimulus condition. Identification accuracy was obtained for conditions in which the two simultaneously-presented stimuli had different amplitude- and/or frequency-modulation contours (extracted from two different SATASK sentences) and for conditions in which the two simultaneously-presented stimuli had identical amplitude- and/or frequency-modulation contours (extracted from a single SATASK sentences). Finally, control data were obtained with the stimuli unmodulated by either amplitude or frequency. These unmodulated stimuli were very similar to the stimuli used by Culling and Summerfield, except that they were of much longer duration.

B. Results and discussion

Figure 3 summarizes the results for the two listeners, plotted separately. Percentage of correct identifications is plotted for each of the four types of stimulus conditions. The lighter bars indicate data obtained when the modulation contours for the two vowels were drawn from different utterances and the darker bars indicate data obtained when the modulation contours of the two whispered vowels were identical.

Because there were only four possible responses, chance performance is nominally 25 percent correct. The fact that the no-modulation condition results in performance near chance confirms the major results of Culling and Summerfield experiment. That is, when there was no modulation of the stimuli, the listeners were unable to use the ITD of 363 µs in order to perceptually segregate and identify the target vowel.

Performance improved substantially when the vowel sounds were modulated by two different frequency and/or amplitude contours, an outcome that is consistent with the results of Experiment 1. For those conditions, identification accuracy was between 54 and 61 percent correct for Subject AR and between 89 and 90 percent for Subject RS. (These differences in performance are likely to be a consequence of RS's having had vast experience as a subject in binaural hearing experiments.) As indicated by the solid bars, Subject RS was also able to perform the discriminations at a level of between 54 and 75 percent correct when the two vowel sounds were modulated by the same pitch and/or amplitude contours, although the identification task was considerably more difficult than it was when the modulating waveforms were different for the two vowels. RS's data are well above chance for all conditions, even assuming that he is able to make use of all available monaural information. Subject AR obtained relatively little improvement in performance over chance when the two vowels were modulated identically, although amplitude modulation provided a small benefit.

III. General conclusions

We conclude from the results of the two experiments that ITD per se can indeed be an extremely useful cue for fostering the identification of simultaneously-presented speech-like sounds. The results of Experiment 1, particularly with whispered speech, indicate that ITDs can facilitate speech intelligibility even when "grouping" cues based on common fundamental frequency are not available. The results of Experiment 2 indicate that listeners are able to use ITDs to foster the separation and identification of speech-like sounds and whispered vowels when the stimuli contain naturally-occurring amplitude and frequency modulations. We note that the information derived from modulating the stimuli is, in principle, monaural in nature and appears to be necessary for the binaural cues based on ITD to become effective in increasing the intelligibility of competing sources of speech. Based on these results, it appears that Culling and Summerfield's findings were at least in part a consequence of the absence of dynamic variations in their stimuli.

ACKNOWLEDGMENTS

This research was partially supported by a graduate research assistantship to Angelo Ripepi from the Carnegie Mellon University Biomedical Engineering Program. C. Trahiotis was supported by Research Grant No. NIH DC-00234 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health. The experimental data reported here were obtained by Angelo Ripepi in partial fulfillment of the requirements for the degree of Master of Science at Carnegie Mellon University.

FIGURE CAPTIONS

Figure 1. The average value of the instantaneous number of coincidences as a simultaneous function of characteristic frequency and internal delay. The stimulus is a low-frequency bandpass noise with a center frequency of 500 Hz and an ITD of –1.5 ms.

Figure 2. Percent correct identification of the content words in the sentences of Experiment 1. Conditions from left to right are natural speech, vocoded speech with natural pitch tracks, vocoded monotone speech with the same fundamental frequency (F0) for target and masker, vocoded monotone speech with different fundamental frequencies for target and masker, and vocoded whispered speech. The lighter bars represent results obtained with target ITD of 363 µs while the darker bars represent data obtained with target ITD of zero µs. Masker ITD was zero ms in all cases.

Figure 3. Results of Experiment 2. Whispered vowels presented with frequency modulation, amplitude modulation, both frequency and amplitude modulation, and no modulation were identified. The lighter bars indicate data obtained when the modulation contours for the two vowels were drawn from the different utterances and the darker bars indicate data obtained when the modulation contours of the two whispered vowels were identical.

REFERENCES

A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, Cambridge, MA, 1990.

Bronkhorst, A. W., and Plomp, R. (1990). “A clinical test for the assessment of binaural speech perception in noise,” *Audiology* **29**, 275–285.

Bronkhorst, A. W., and Plomp, R. (1992). “Effect of multiple speech-like maskers on binaural speech recognition in normal and impaired hearing,” *J. Acoust. Soc. Am.* **92**, 3132-3139.

Culling, J. F., and Summerfield, Q. (1995). “Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay,” *J. Acoust. Soc. Am.*, **98**, 785-797.

Darwin, C. J., and Carlyon, R. P. (1995). “Auditory Grouping,” in *Handbook of Perception and Cognition, Volume 6: Hearing*, edited by B. C. J. Moore (New York: Academic Press),

Darwin, C. J., and Hukin, R. W. (1997). “Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity,” *J. Acoust. Soc. Am.* **102**, 2316–2324.

Hawley, M. L., Litovsky, R. Y., and Colburn, H. S. (1999). “Speech intelligibility and localization in a multi-source environment,” *J. Acoust. Soc. Am.* **105**, 3436-3448.

Hukin, R. W., and Darwin, C. J. (1995). “Effects of contralateral presentation and of interaural time differences in segregating a harmonic from a vowel,” *J. Acoust. Soc. Am.* **98**, 1380–1387.

Koehnke, J., and Besing, J. M. (1996). “A Procedure for Testing Speech Intelligibility in a Virtual Listening Environment,” *Ear & Hearing*,**17**, 211-217.

Nilsson, M. J., and Soli, S. D. (1994). “Norms for a headphone simulation of the hearing in noise test: comparison of physical and simulated spatial separation of sound sources,” *J. Acoust. Soc. Am.* **95**, 2994.

Oppenheim, A. V., Schafer, R. .W., and Buck, J. R. (1999). *Discrete-Time Signal Processing (Second Edition),* Upper Saddle River, NJ: Pearson Education Publishers.

Peissig, J., and Kollmeier, B. (1997). '"Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners," *J. Acoust. Soc. Am.* **101**, 1660–1670.

Rabiner, L. R., and Schafer, R. W. (1978). *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs NJ.

Talkin, D. (1995), "A Robust Algorithm for Pitch Tracking (RAPT)", in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., Amsterdam, NL: Elsevier Science, pp. 495-518.

Yost, W. A. (1992). "Auditory Image Perception and Analysis," *Hearing Res.* **56**, 8-19.

Yost, W. A., Dye, R. H. Jr., and Sheft, S. (1996). "A Simulated "Cocktail Party" with Up to Three Sound Sources," *Percept. Psychophys.* **58**, 1026–1036.

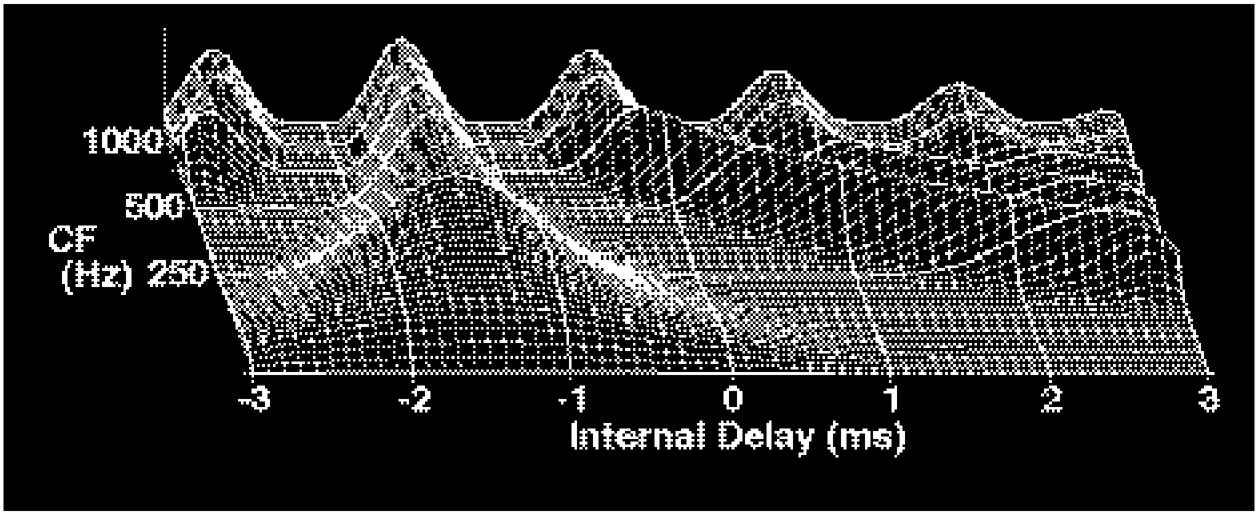

Figure 1. The average value of the instantaneous number of coincidences as a simultaneous function of characteristic frequency and internal delay. The stimulus is a low-frequency bandpass noise with a center frequency of 500 Hz and an ITD of –1.5 ms.

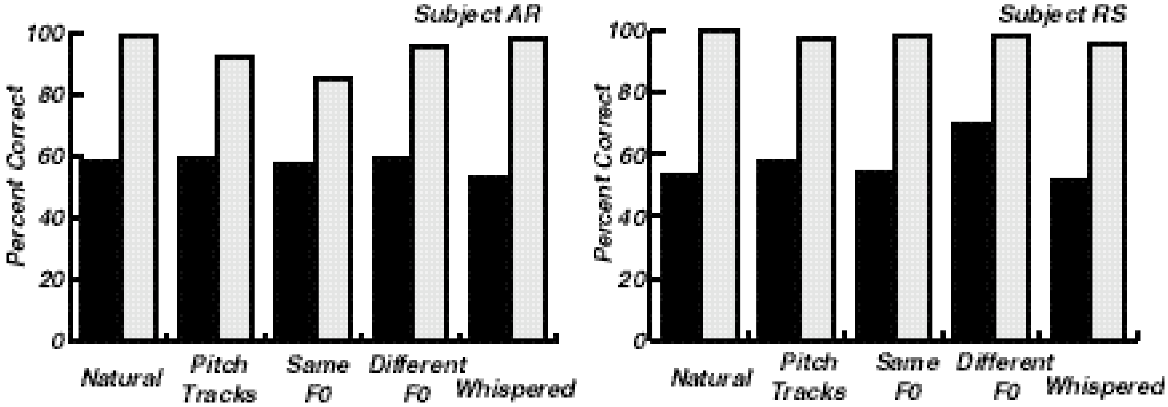

Figure 2. Percent correct identification of the content words in the sentences of Experiment 1. Conditions from left to right are natural speech, vocoded speech with natural pitch tracks, vocoded monotone speech with the same fundamental frequency (F0), vocoded monotone speech with different fundamental frequencies, and vocoded whispered speech. The darker bars represent results obtained with target ITD of zero µs while the lighter bars represent data obtained with target ITD of 363 µs. Masker ITD was zero ms in all cases.

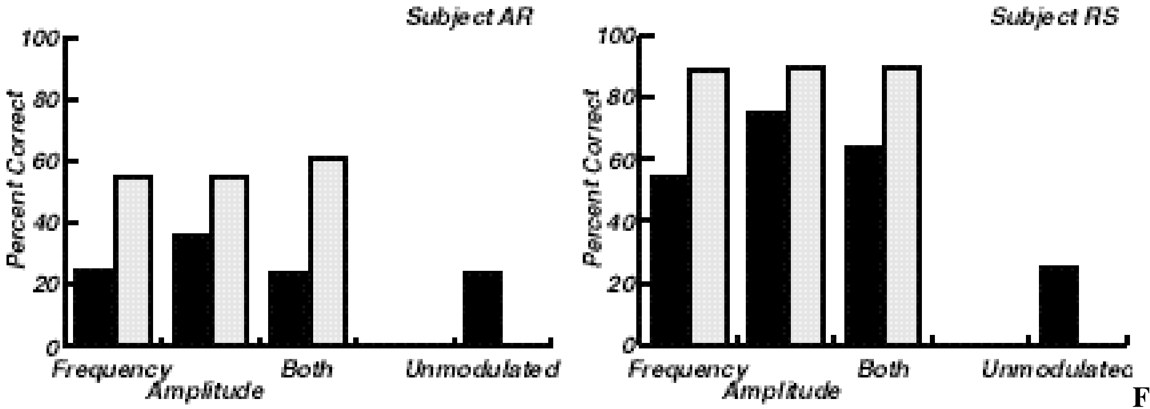

Figure 3.

Figure 3. Results of Experiment 2 whispered vowels presented with frequency modulation, amplitude modulation, both frequency and amplitude modulation, and no modulation were identified. The lighter bars indicate data obtained when the modulation contours for the two vowels were drawn from the different utterances and the darker bars indicate data obtained when the modulation contours of the two whispered vowels were identical.