# LATERALIZATION AND IDENTIFICATION OF SIMULTANEOUSLY-PRESENTED WHISPERED VOWELS AND SPEECH SOUNDS

**Angelo M. Ripepi**

Submitted to the Department of Electrical and Computer Engineering in Partial Fulfillment of the Requirements for the Degree of Master of Science in Electrical and Computer Engineering (Bioengineering)

Carnegie Mellon University
Pittsburgh, PA 15213

May 14, 1999

# Abstract

This report describes the results of two different sets of experiments that investigate the extent to which one can identify competing speech-like or vowel signals solely on the basis of their interaural time delay (ITD), the primary cue available for auditory lateralization at low frequencies. The first experiment demonstrated that human subjects can use ITD cues to separate and identify competing natural speech and degraded speech sentences, even when there is no consistent information based on fundamental frequency available to discriminate the signals. The subjects ability to separate degraded speech showed the power of ITD cues. The second experiment explored human s inability to separate competing whispered vowel sounds with amplitude and/or frequency modulation derived from natural speech sounds. The resulting contoured vowel sounds could be separated with interaural timing cues when the applied amplitude contours or pitch contours were different for competing vowel sounds. When competing vowel sounds used the same amplitude or pitch contours, the ability to separately identify the signals on the basis of interaural timing cues diminished significantly. The results of these experiments suggest that ITD cues are useful in separating and identifying simultaneously-presented speech-like signals and vowel sounds, but the ease with which this can be done is easily disrupted when similar amplitude and especially similar frequency contours are applied to the competing signals.

# 1. Introduction

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 (*e.g.* Hukin and Darwin 1994, Koehnke and Besing 1996). These observations are confirmed by results of many quantitative perceptual mesurements of word recognition accuracy in the presence of noise. (*e.g* Bronkhorst and Plomp, 1992).

The two major physical cues that arise from spatial separation of sound sources are interaural time differences (ITDs) and interaural intensitive differences (IIDs) (*e.g.* Shaw 1997), with information based on interaural timing differences being especially salient in separating sound sources at the most important frequencies in the speech waveform.

Most currently-popular models of binaural interaction assume that sound impinging on the two ears is first subjected to peripheral frequency analysis in the cochlea that can be modeled as a bank of bandpass filters with differing center frequencies. Information pertaining to interaural timing differences is presumed to be extracted by a putative neural coincidence mechanism that is assumed to compare the outputs of similar bandpass-frequency channels in the auditory periphery (Stern and Trahiotis, 1994, 1997). The output of this mechanism closely resembles the interaural cross-correlation function of the stimuli after the peripheral bandpass filtering operation. It is frequently convenient to think of the outputs of this binaural display mechanism as a function of two variables, the stimulus frequency and the internal delay parameter of the cross-correlation function.

For many years, researchers had assumed that people could separately localize and perceive simultaneously-presented sounds coming from different directions (and hence arriving at the two

ears with differing ITDs) by focussing on a single set of internal delays of the binaural processing mechanism (*e.g.* Cherry, 1953; Cherry and Taylor 1954). This view remained largely unchallenged for decades, as results of many experiments indicated that simultaneously-presented speech sounds, for example, are more intelligible when they are presented with different ITDs (*e.g.* Rayleigh, 1907; Jeffress, 1948; Dirks and Wilson, 1969; Koehnke and Besing; 1996).

Nevertheless, the results of a recent study by Culling and Summerfield (1995) indicated that simultaneously-presented artificially-generated vowel sounds cannot be separately identified even when presented with large differences in ITDs. The purpose of the present study is to attempt to understand some of the reasons underlying the differences between the results of Culling and Summerfield and those of many studies that indicate that natural speech sounds are easily separated on the basis of their ITDs.

In Chapter 2 we will review some of the background and previous work leading up to this study. In Chapter 3 we will present the results of a first series of experiments that were developed to determine the extent to which the inability to separately-identify simultaneously presented vowels also extends to more speech-like stimuli. In Chapter 4 we present the results of a complementary experiment that explores people s ability to identify simultaneously-presented vowel sounds presented with amplitude and frequency modulation. Chapter 5 contains a discussion and summary of the findings.

## 2. Background and Previous Work

As noted in the previous chapter, this project is motivated by certain observations of Culling and Summerfield (1995) concerning subjects inability to identify vowel sounds that are presented simultaneously, but with differing ITDs.

Culling and Summerfield created a set of whispered vowel sounds that are produced by passing broadband noise through a pair of bandpass filters that approximate the first two formant frequencies of common vowels. The bandpass filters had non-overlapping passbands with center frequencies, 225, 625, 975, and 1925 Hz. These frequencies were chosen because when one of the two lower bands was presented in combination with one of the two higher bands a recognizable vowel sound ( AR , EE , OO , or ER ) was produced. Example spectrograms for these sounds are shown in Figure 1 below.

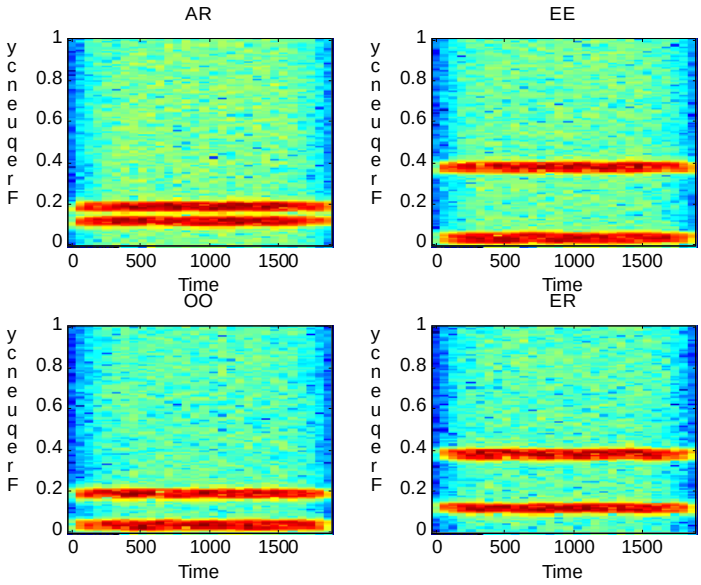


**Figure 1.** Spectrograms of Culling and Summerfield s whispered vowels . The frequency plots are vertically normalized such that the maximum frequency value ( 1 ) corresponds to the Nyquist frequency, 5512.5 Hz.

While there were a number of experiments included in the Culling and Summerfield (1995) study, the result most relevant to the present work was that if pairs of these vowels were

presented simultaneously with non-overlapping frequencies (which only occurs if AR is presented together with EE or OO is presented with ER), listeners cannot identify which vowels are presented, even if the two vowel sounds are presented with different ITDs. This observation is contrary to what would normally be expected from models such as those reviewed by Stern and Trahiotis (1995), which imply that one can separately perceive the two vowel sounds by focussing attention at the internal delay along the internal cross-correlation display produced by the model that corresponds to the ITD of the stimulus component.

Culling and Summerfield believe that their observations are a reflection of the grouping that the auditory system performs on components of complex sounds. Auditory grouping (sometimes called auditory scene analysis ) is the process by which the auditory system clusters components of a complex sound field that are believed to arise from a common sound source. The processes mediating auditory grouping have been studied by psychologists for decades and are extensively reviewed in Bregman (1990). Some attributes of components of sounds that are believed to contribute to their being grouped together include common fundamental frequency, common onset and/or offset, common amplitude or frequency modulation, and common ITD.

Culling and Summerfield (1995) believe their results indicate that lateralization cues such as ITD can only be used to identify components of simultaneously-presented sound sources only after they are grouped by the auditory system. (In contrast, the cross-correlation-based models imply that the ability to separate sources by ITD is inherent, and not dependent on the components being pre-grouped by the auditory system. Specifically, Culling and Summerfield observe that listeners could not group the noisebands in different frequency regions with the same ITD and thereby separate them from bands in other frequency regions with a different ITD . They also suggest that for the purpose of identifying speech sounds in noise, the auditory system exploits interaural timing differences between speech and noise within each frequency channel

independently. Culling and Summerfield go on to suggest, the analysis ignores correspondencies between the interaural delays in different frequency channels .

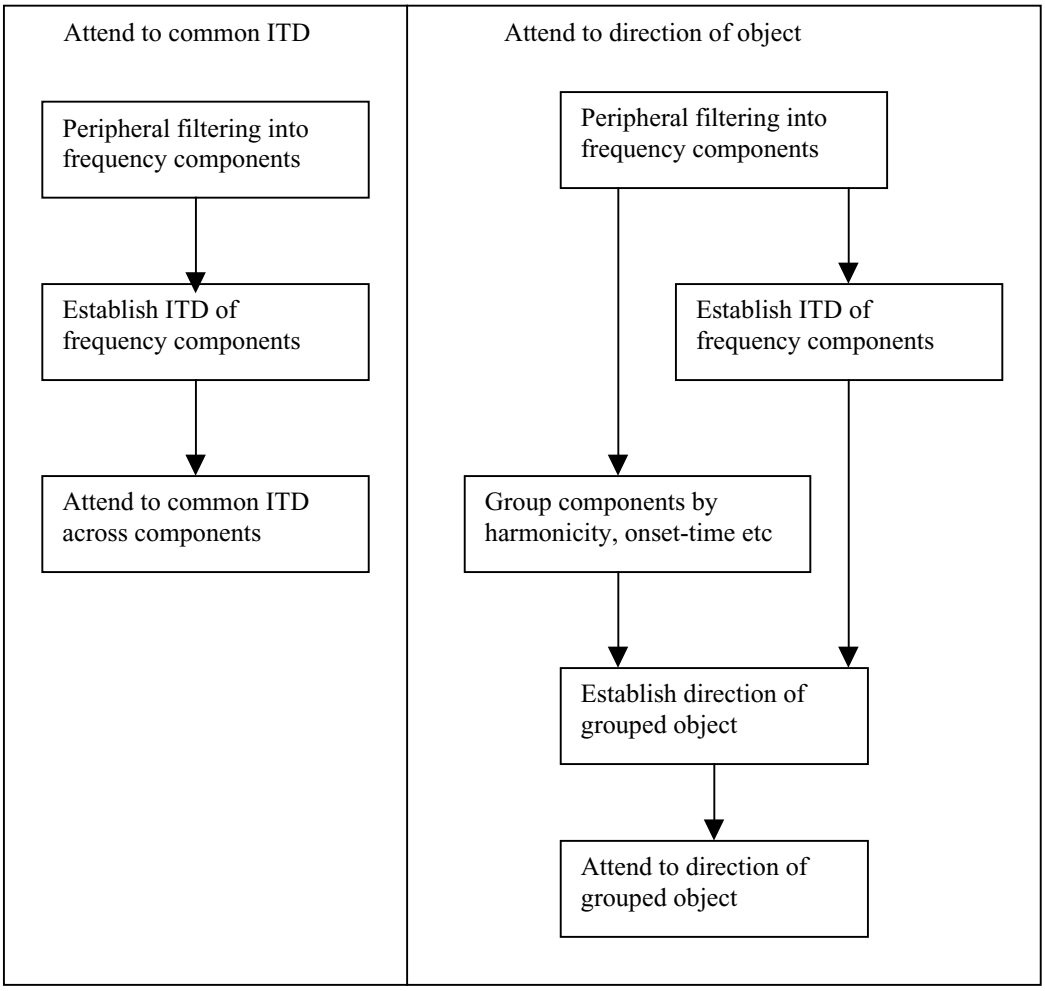


**Figure 2.** Schematic comparison of theories that assume that extraction and availability of ITD cues is inherent (left panel) with theories that postulate that grouping must precede separation by ITD (right panel).

This view is shared by some other researchers, including Darwin and Hukin (1998), who have proposed the schematic model shown in Figure 2. for understanding the relationship between separation by ITD and grouping.

The results of the Culling and Summerfield experiments are counterintuitive to some researchers not only because they are not predicted by cross-correlation models but also because they are

inconsistent with the common observation that speech perception is improved when targets and maskers are presented with different ITDs. The goal of the experiments in this project is to better understand the extent to which grouping of stimulus components is necessary for lateralization cues such as ITD to be useful in separately perceiving components of sounds coming from different locations. This is done using two sets of stimuli that are intermediate in nature between the whispered vowels used by Culling and Summerfield and natural speech sounds. In Chapter 3 we describe the results of a series of experiments in which potentially useful grouping information is removed from natural speech sounds by eliminating pitch and harmonicity information. In Chapter 4 we describe results obtained when the whispered vowels of Culling and Sumerfield are rendered more identifiable by the addition of amplitude and frequency modulation.

# 3. Identification of simultaneously-presented natural and degraded speech

In this chapter we describe the results of a series of experiments intended to determine the extent to which subjects can identify messages presented binaurally with different ITDs. Experiments were performed using both natural speech, and speech with some of the information that can lead to grouping removed.

In an attempt to understand Culling and Summerfield s results natural and degraded speech was presented to the subjects. The speech was presented to the subject in the form of a target and a masking sentence. The sentences were spatially separated with varying ITD s. The subject would attempt to focus on the target sentence while the masking sentence was played simultaneously. The target sentence would always start with the word TROY .

## *3.1 Stimuli and Experimental Procedure*

The stimuli used for these for these experiments were taken from the SATASK database recorded by the US Army Research Laboratory (Koehnke and Besing 1996). These sentences were obtained from the Speech Pathology and Audiology Department at the University of South Alabama in Mobile, AL, with the help of Janet Koehnke and Joan Besing. The sentences were originally sampled at 22050 Hz and then downsampled to 11025 Hz for the present experiments. The sentences were recorded by each of four male speakers with the following carrier phrase: (NAME) write the number (NUMBER) on the (COLOR) (OBJECT) . The sentences were recorded under carefully controlled conditions to ensure, to the extent possible, that when multiple utterances are time aligned, the major content words of each of the utterances are spoken simultaneously.

The sentences were presented in four different fashions: natural, monotone, monotone with varying pitch, and whispered. Natural speech is simply the original speech as recorded in the SATASK database. The monotone and whispered speech was obtained using LPC waveform coding methods. The incoming speech was windowed using 20-ms Hamming windows, overlapped by 10 ms. Fourteen LPC coefficients were obtained for each windowed segment using the Levinson-Durbin method (Rabiner and Schafer, 1978). These coefficients characterize the time-varying spectral profile of the incoming speech, but without detailed information about the excitation signal. The monotone speech signals were obtained by exciting the time-varying LPC coefficients with a periodic impulse train. The monotone speech with varying pitch was obtained by exciting the LPC coefficients with an impulse train, but with an instantaneous frequency that equalled the fundamental frequency of the voiced segments of the original signal, as estimated using the pitch algorithm of the commercial Entropic Signal Processing System (ESPS) by Entropic Research Laboratory 1991. The whispered speech was obtained by exciting the LPC coefficients with white noise. The whispered speech was obtained by exciting the LPC coefficients with white noise. The use of the monotone and whispered speech signals enable us to present speech-like stimuli for which the grouping cues associated with fundamental frequency are either difficult to separate (as in the case of the monotone speech presented with the same fundamental frequency), or non-existent (as in the case of the whispered speech).

A total of 9216 possible sentences were available for the experiments composed of the items listed in Table 1.

| NAME | NUMBER | COLOR | OBJECT |
|---|---|---|---|
| Troy | 1 | Red | Ball |
| Mike | 2 | Blue | Cup |
| Nate | 3 | Green | Fork |
| Ron | 4 | Pink | Key |
| | 5 | Brown | Kite |
| | 6 | Black | Spoon |
| | 8 | White | Square |
| | 9 | Gray | Stair |
| | | | Star |

**Table 1.** Stimulus combinations used in the sentence experiments.

Pairs of sentences from the SATASK database were combined digitally and presented binaurally. One of the two sentences would always begin with the NAME Troy and is referred to as the target sentence, and the other sentence is referred to as the masker sentence. 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 same position in the center of the head when presented dichotically using 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 400- s ITD, causing the latter sentence to be perceived much closer to the right ear.

Pairs of sentences would be presented to the subject using headphones in a soundproof room, typically in blocks of 25 trials. The subject would attempt to discriminate between the target sentence and the masking sentence by listening for the NAME Troy . The pairs of sentences presented always used speech from two different speakers, and the NAMEs, NUMBERs, COLORs, and OBJECTs used within each target-masker set were unique. An example of a target

sentence would 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 .  Using the graphical user interface (GUI) shown below the subject would choose the NUMBER, COLOR, and OBJECT that corresponded to the sentence that started with the name TROY.

The subject responded by clicking on the appropriate radio buttons, and then he or she clicked the NEXT button to initiate the next set of sentences.  If the subject omited a response in using the GUI,  a message was displayed in the box at the top of the screen.  The results were compiled and the percentage of correct responses were tabulated for the total number of responses and on an item-by-item basis.

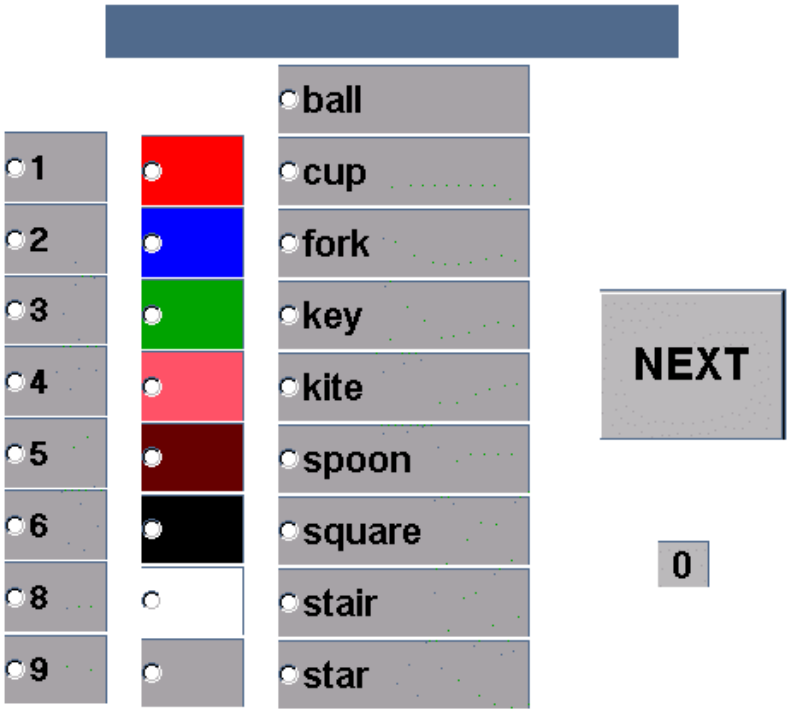


**Figure 3.** Graphical User Interface developed for the sentence identification experiments.

Experiments were conducted using natural speech, monotone speech in which both sentences were presented with the same fundamental frequency (typically 100 Hz), monotone speech in

which the sentences were presented with two different fundamental frequencies (typically 100 and 200 Hz), LPC-derived speech that was excited by a pulse train with instantaneous frequency that tracked the fundamental frequency of the original sentences, and whispered speech (in which both sentences were presented with statistically independent noise excitations. The sentences were presented to two subjects, each subject going through four blocks of 25 sentence pairs for each set of stimuli. In other words, results were based on 100 sentences from each subject for each stimulus condition.

### *3.2 Experimental Results and Discussion*

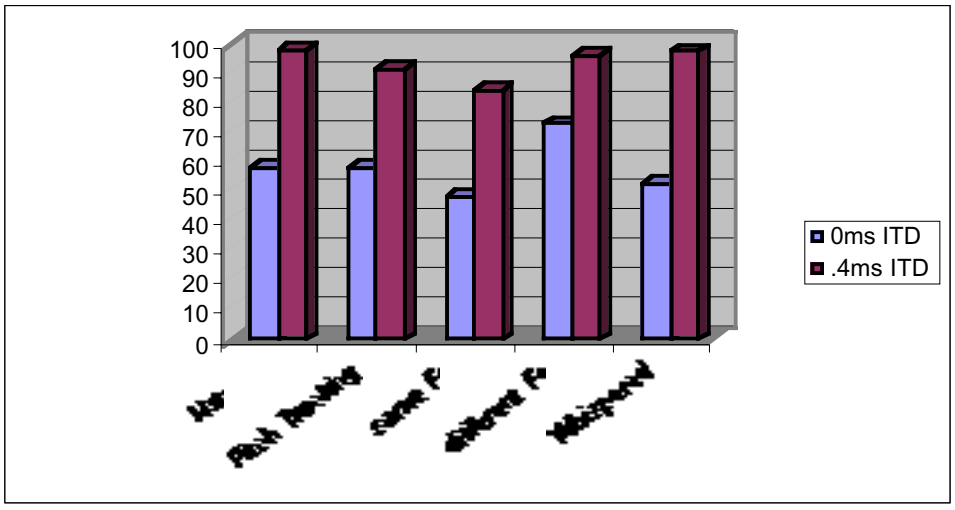


**Figure 4.** Recognition accuracy obtained for the content words in the degraded sentences, averaged over two subjects.

Results for the experiments using the simultaneously-presented speech-like sounds are summarized in Figure 4, averaged over the two subjects. In all cases, the masker sentence was presented with zero ITD. Results are plotted separately for two different target ITDs, 0 and 400 s.

With 8 numbers, 8 colors, and 9 objects to identify, chance performance for this task is about 12 percent. The actual percentage correct obtained by the two subjects was substantially above chance for all conditions and for both ITSs. One probable contributing factor to the ability to achieve identification performance that is much better than chance even when both target and masker sentences are presented with zero ITD is the incompleteness of the masking when the masker consists of a single sentence. (We also used maskers consisting of multiple sentences in some of the pilot experiments, but here distinguishing the target from the masker at 0-dB target-to-masker ration was relatively easy, since the target sentence clearly emerged from the background babble. Nevertheless, it is also clear that identification performance is much better when the sentences are presented with different ITDs, indicating that interaural timing differences remain a useful cue at least for identifying these stimuli. In fact, there were surprisingly little differences in the nature of the results for the various stimulus conditions. Not surprisingly, the most difficult condition appeared to be monotone sentence pairs with the same fundamental frequencies, but for even these sentences overall performance was well above chance, and identification accuracy was much greater when the two sentences were presented with differing ITDs.

We conclude from these observations that ITD is an extremely useful cue that substantially improves identification accuracy for simultaneously-presented speech-like sounds. Although Culling and Summerfield (1995) had argued that grouping was needed for the signals to be identified separately, these results, particularly with whispered speech, indicate that identification by lateralization can be possible, even when grouping cues based on common fundamental frequency are not available. Nevertheless, even these results cannot rule out the possibility that the high level of observed identification accuracy was facilitated by grouping of stimulus components according to common amplitude or frequency modulation. We explore this latter possibility with the second set of experiments described in the following chapter.

## *4. Identification of simultaneously-presented vowel sounds with speech-like pitch and amplitude contours*

This chapter describes the results of a complementary set of experiments to evaluate the extent to which the auditory lateralization mechanism provides useful information in processing speech-like sounds, even in the absence of auditory grouping. As in the previous chapter, we make use of stimuli that are intermediate in nature between natural speech and the whispered vowels used by Culling and Summerfield. While in Chapter 3 we generated signals using by degrading natural speech sounds. In this chapter we work in the opposite direction, beginning with the whispered vowels of Culling and Summerfield, and then adding amplitude and frequency information to render them more like speech.

### *4.1 Stimuli and Experimental Procedures*

The stimuli used in this series of experiments consisted of our own local implementation of the Culling and Summerfield whispered vowels and versions of these signals that were amplitude modulated and/or frequency modulated by the amplitude and frequency contours of natural speech.

Realizations of the whispered vowels were obtained by passing white noise through (time-invariant) filters with two narrow rectangular passbands, as used in the original Culling and Summerfield experiemtents. Typically these filters were finite impulse response (FIR) equiripple filters, with the four center frequencies used in the original experiments (225, 625 975, and 1925 Hz), 50-Hz transitional bands, and a filter length of 512. Each different signal was produced by exciting the bandpass filter with a statistically-independent excitation function.

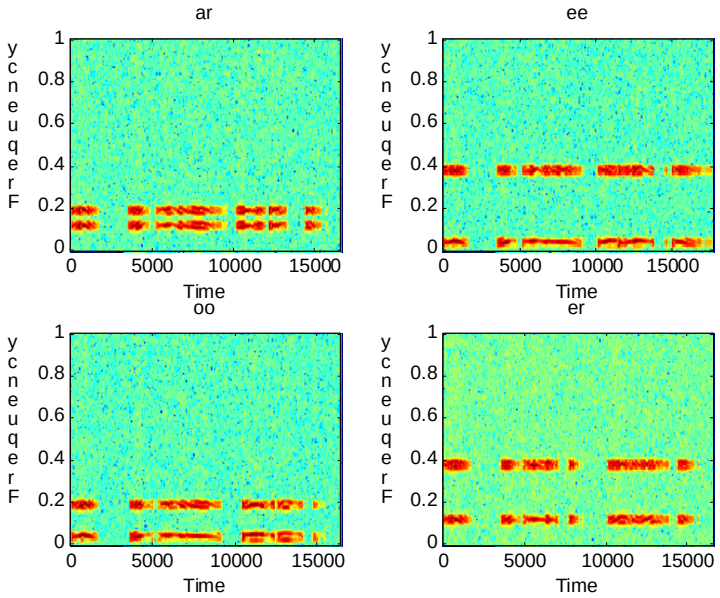


**Figure 5.** Examples of whispered vowels with amplitude and frequency modulation.

Amplitude-modulation contours were obtained by measuring the short-term energy of speech waveforms from Koehnke and Besing s SATASK database, selected as they were in the experiments described in the previous chapter. Frequency-modulation contours were obtained by estimating the pitch of the SATASK sentences, again using the commercial ESPS package by Entropic Research Laboratory. Figure 5 shows examples of spectrograms of SATASK sentences presented with both amplitude and frequency modulation.

Experiments were conducted using the unmodulated whispered vowel stimuli, the whispered vowels with natural amplitude-modulation contours, the whispered vowels with natural frequency-modulation contours, and the whispered vowels presented with both amplitude and frequency modulation. Gaussian noise, at +30-dB SNR, was added to the contoured vowel sounds to mask components at frequencies other than the vowel sound frequencies. The unmodulated whispered vowel sounds were similar to the stimuli used by Culling and Summerfield, expect that our signals were much longer in duration.

For each block of trials, two vowel waveforms, either an AR and an EE or an OO and an ER were added digitally at the 0-dB target to masker ratio. One of the two vowels was presented with zero ITD and was perceived in the center of the head; the other was presented with a 400- s ITD, and was perceived toward the right ear. The task of the subjects was to identify whether an AR , EE , ER , or OO was presented toward the right side of the head. Subjects knew that only four of the twelve vowel possible combinations of vowel identity and position were actually presented, and in some cases they used this information to improve their identification accuracy.

Experiments were conducted in four blocks of 25 trials for each stimulus condition, with two subjects. Performance was compared for identification when the two simultaneously-presented stimuli had amplitude- and/or frequency-modulation contours extracted from two different SATASK sentences and when the contours for the two vowels were extracted from the same SATASK sentence, producing identical modulation contours.

### *4.3 Experimental Results and Discussion*

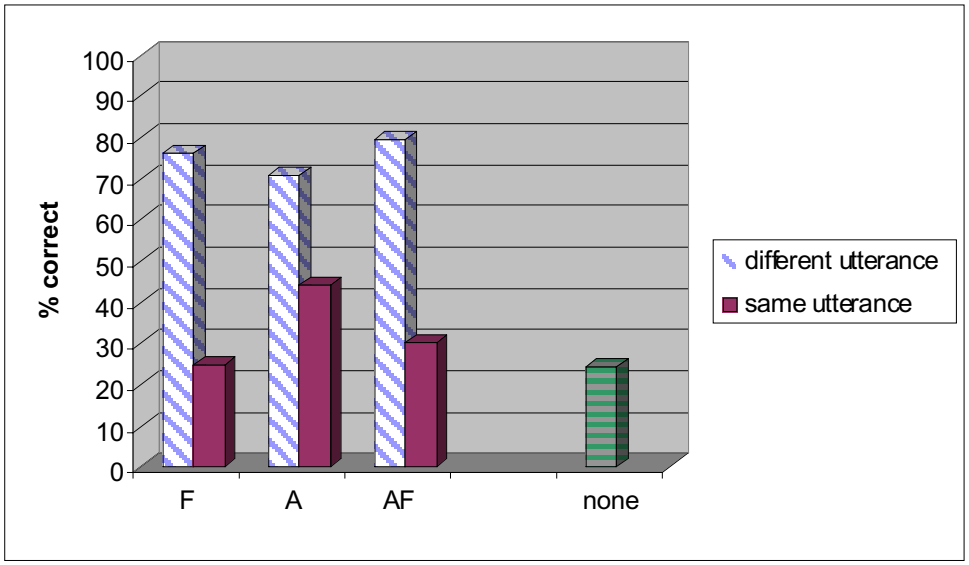


**Figure 6.** Results of the experiments using whispered vowels presented with no modulation ( none ), frequency modulation ( F ), amplitude modulation ( A ), and both amplitude and frequency modulation ( AF ). Results are shown for one subject.

Figure 6 summarizes the results obtained for one subject in the vowel identification experiments. Since there are only four possible responses, chance performance is 25 percent correct. As noted above, the unmodulated vowels (indicated by none in Figure 6) are very similar to the stimuli used by Culling and Summerfield, but longer in duration. Our results confirm those of Culling and Summerfield: even with a 400- s ITD, it is not possible to identify which vowel is presented toward the right ear without modulation.

As can be seen by the results in Figure 6, the task was considerably easier when speech-like amplitude or frequency modulation contours from two different utterances were applied to the vowel sounds. (The second subject achieved near-perfect identification for these stimuli.) Nevertheless, it proved to be much more difficult to identify the vowels when the amplitude- or frequency-modulation contours were drawn from the same SATASK sentence. The subject depicted in Figure 6 scored 44% correct on the amplitude tracking (A) condition, 24.85% on the pitch tracking (F) condition, and 30% on the amplitude and pitch (AF) tracking condition. The performance of the second subject, for whom quantitative data are incomplete, was somewhat better, but still exhibited the same qualitative trends: overall identification accuracy was much worse than when the competing sounds were modulated by different amplitude or pitch contours, performance with amplitude modulation only was better than with frequency modulation only, and performance with both amplitude and frequency modulation was intermediate.

We consider these results to be consistent with our findings in Chapter 3. While speech-like signals and modulated whispered vowels can be identified on the basis of ITD, the ease with which identification can take place is strongly affected by other factors, and most strongly by the presence of differing amounts of amplitude modulation. These findings offer partial support to the hypothesis of Culling and Summerfield. Specifically, information relevant to the perceptual

grouping of the stimuli strongly affects the ease with which signals can be identified on the basis of ITD. Nevertheless, not all grouping cues are equally potent, with frequency contours playing a much more significant role in inhibiting identification accuracy than amplitude contours. Furthermore, identification performance that is significantly above chance can still take place on the basis of ITD information alone, and in the absence of other cues that are useful for grouping, although the task is much more difficult.

## 5. General Discussion and Summary

Previous experimental results, as well as those obtained in this study, have shown that ITD is a very powerful tool when it comes to spatially separating competing speech. In our experiments simultaneously-presented degraded speech sounds were also easily separated by ITD. Culling and Summerfield have proposed and we have confirmed that ITD alone is not adequate to separate whispered vowel sounds. Nevertheless, whispered vowels can be easily separated and identified if they are presented with differing pitch and/or amplitude contours, as are obtained from separate natural speech utterances. Separation and identification of speech-contoured whispered vowel sounds became much more difficult when the speech contours applied to the competing vowel sounds came from the same utterance, and separate identification was totally impossible if only identical frequency modulation is applied to the stimuli.

Among cues based on amplitude modulation, frequency modulation, and interaural timing information through ITDs, the pitch-tracking cues appear to be the most potent in that the vowel sounds cannot be separately identified (even when presented with different ITDs) if they are presented with the same fundamental frequency contours. However, ITD cues are more salient than amplitude-tracking cues because when whispered vowel sounds are presented with identical amplitude-tracking cues the competing sounds are still separately identifiable.

## *References*

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

A. W. Bronkhorst and R. Plomp, Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing, *J. Acoust. Soc. Amer.* Vol. 92, pp. 3132-9, 1992.

E. C. Cherry, Some experiments on the recognition of speech with one and with two ears, *J. Acoust. Soc. Amer.*, Vol. 25, pp. 975-979, 1953.

E. C. Cherry and W. K. Taylor, Some further experiments upon recognition of speech, with one and two ears, J. Acoust. Soc. Amer. Vol. 25, 975-979, 1954.

J. F. Culling, C. J. Darwin, Perceptual and computational separation of simultaneous vowels: cues arising from low-frequency beating, *J. Acoust. Soc. Amer.*, Vol. 95, pp. 1559-1569, 1994.

J. F. Culling, Q. Summerfield, Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay, *J. Acoust. Soc. Amer.*, Vol. 98, pp. 785-797, 1995.

C. J. Darwin, R.W. Hukin, Auditory Objects of Attention, *Proc. of the 135th Meeting of the Acoustical Society of America*, pp. 1571-1572, 1998.

D.D. Dirks, R.A. Wilson, The effect of spatially separated sound sources on speech intelligibility, *Journal of Speech and Hearing Research* , Vol. 12, pp. 5-38, 1969.

L.A. Jeffress, A place theory of sound localization, *Journal of Comparative and Physiological Psychology*, Vol. 41, pp. 35-39, 1948.

D.M. Green, *An Introduction to Hearing*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1976.

J. Koehnke, J.M. Besing , A Procedure for Testing Speech Intelligibility in a Virtual Listening Environment, *Ear & Hearing* , Vol. 17, No. 3, pp. 211-217, 1996.

L. Rayleigh, On our perception of sound direction, *Philosophical Magazine* , Vol. 13, pp. 214-232, 1907.

E. A. G. Shaw, Acoustical Features of the Human External Ear, In *Binaural and spatial hearing in real and virtual environments*, R.H. Gilkey, T.R. Anderson, Editors, Lawrence Erlbaum Associates, Mahwah, NJ, 1997.

R. M. Stern and C. Trahiotis, Models of Binaural Interaction, *Hearing* , B.C.J. Moore, Editor, Academic Press, New York, NY, Vol. 6, pp. 347-386, 1994.

R. M. Stern and C. Trahiotis, Models of Binaural Interaction , In *Binaural and spatial hearing in real and virtual environments*, R.H. Gilkey, T.R. Anderson, Editors, Lawrence Erlbaum Associates, Mahwah, NJ, 1997.