THE DEVELOPMENT OF AMPLITUDE MODULATION AS AN AUDITORY GROUPING CUE

By

Dawna E. Lewis

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Interdepartmental Area of Human Sciences

Under the Supervision of Professor Thomas D. Carrell

Lincoln, NE

May, 2005

DISSERTATION TITLE

The Development of Amplitude Modulation as an Auditory Grouping

BY

Dawna E. Lewis

SUPERVISORY COMMITTEE:

Approved	Date
Thomas Carrell	4-15-05
Signature	
Thomas Carrell, Ph.D.	
Typed Name	
T W Decker	4/15/05
Signature	
Newell Decker, Ph.D.	
Typed Name	
Dixie Sanger	4/15/05
Signature	
Dixie Sanger, Ph.D.	
Typed Name	
John B Flowers	4/15/05
Signature	
John Flowers, Ph.D.	
Typed Name	
Signature	
Typed Name	
Signature	
Typed Name	

University of Nebraska Lincoln

THE DEVELOPMENT OF AMPLITUDE MODULATION AS AN AUDITORY GROUPING CUE

Dawna E. Lewis, Ph.D.

University of Nebraska, 2005

Adviser: Thomas Carrell

The purpose of this research project was to determine whether children use amplitude modulation (AM) as a grouping cue in the formation of auditory objects. Children from 4 to 13 years of age, as well as adults, were evaluated to examine whether children's ability to use AM in sentence perception is different from adults and whether this ability changes with development. Although researchers are currently studying auditory object formation in adults, little is known about the development of this phenomenon in children. Because the study of auditory grouping in speech perception is complicated by the fact that speech contains numerous redundant acoustic characteristics, time varying sinusoidal (TVS) speech was used. New stimuli were developed based on age-appropriate sentences. Performance on recognition of TVS sentences (unmodulated, amplitude comodulated at four frequencies, and amplitude modulated using conflicting frequencies) was evaluated. In general, the youngest children performed more poorly on recognition tasks than did older children and adults. However, accuracy improvements based on amplitude modulation showed no significant effects of age or language level. Difference scores, defined as percent phonemes correct in a given modulation condition minus percent correct for the unmodulated (0mod) condition, indicated that subjects performed more poorly in the 200 Hz modulation condition than they did in the unmodulated condition.

While difference scores were positive for all other modulation conditions, low difference scores for the 25 Hz modulation condition compared to 50 Hz, 100 Hz, and conflicting modulation conditions would seem to preclude a CMR-like mechanism underlying performance. While previous research had suggested that amplitude *comodulation* was necessary for improving sentence recognition, subjects in the present study performed as well with conflicting amplitude modulation as with amplitude comodulation at 100 Hz. While the this study provides evidence that children and adults receive the same benefits (or decrements) from amplitude modulation, the underlying causes for listeners' performance remain unclear. Additional research will be needed to investigate those causes. (This research was supported by NIH grants F31 DC006582 and P30 DC04662)

ACKNOWLEDGMENTS

I am humbled and amazed at the many people who have been a part of this journey with me. First, I must thank Randy, who is an amazing husband and father and who has been my rock during the past five years (and beyond). Without your support and encouragement I would never have gotten through all of the ups and downs I encountered. I can never repay you for all of the responsibilities you shouldered ("carpool dad", "homework helper", and consummate multi-tasker), and for the patience you displayed, especially when I felt overwhelmed. Know that you are my hero and I will always love you.

I also thank Jessica and Jaimie. As daughters, I could not have asked for any better. You never made me feel guilty or complained about the time that school required. That was a wonderful gift. You have grown into wonderful young women.

I thank my sister Karen for all of the help over the past five years. Having you with us, helping in so many ways with the daily needs of our family was a real blessing. I'm sure I haven't told you often enough how much I appreciate what you have done.

I am thankful for my parents, Charles and Elizabeth Johnson, who have always loved me and supported me. You are a big reason I am the person I am today. You instilled in me the knowledge that I could accomplish much if I worked hard and persevered. I am honored to be your daughter.

There are many people at UNL who also deserve credit. As an advisor and mentor, Tom Carrell has been outstanding and wonderfully supportive right from the start. Tom, you have gone above and beyond to help me with many time-consuming projects—most notably, the grant submission that wouldn't end—and I thank you. I look forward to many more opportunities to collaborate with you in the future.

The faculty and staff at Barkley have all been amazing. I have always felt encouraged and knew that everyone truly desired my success as a doctoral student. I appreciate you all. I am especially appreciative of the support and friendship I received from Jody Spalding. As I fumbled my way around supervising brand new graduate students, Jody, you were always open and supportive. I gained not only a colleague but a friend. Thank you.

Next, I must thank my coworkers at BTNRH. Pat, you have been an amazing boss for the past 23 years. Without your support and willingness to be so flexible, I never would have been able to manage both work and school. Your leadership and example were a huge factor in my decision to pursue this degree. I know I still have much to learn from you and look forward to continued collaboration. Thanks also my coworkers and friends, especially Leisha Eiten, Judy Feigin, Kathy Beauchaine, and Debbie Smith, who have encouraged and supported me through this long journey. You all have been an indispensible part of my success.

Finally, and most importantly, I thank God who has been my rock and my fortress for so many years. As always, You have seen me through another phase of my life in Your amazing and sometimes mysterious way. I could not have predicted what these five years brought but You knew and had a plan even when I didn't have a clue. I know I shouldn't have been surprised. You have always kept your promises and I feel so blessed.

"'For I know the plans I have for you', declares the Lord, 'plans to prosper you and not to harm you, plans to give you a hope and a future'." Jeremiah 29:11

TABLES

FIGURE CAPTIONS

Figure 1. Narrowband spectrogram of the TVS sentence, "The yellow lion roared."

Figure 2. Narrowband spectrograms of the sentence "The swan dive was far short of perfect" representing natural speech, unmodulated TVS speech, and TVS speech comodulated at 100 Hz.

Figure 3. Mean intelligibility scores for UTVS and AMTVS sentences as a function of age group.

Figure 4. Schematic of sawtooth modulation used to modulate the UTVS waveforms.

Figure 5. Wideband spectrograms of the sentence "My tooth is loose" for each of the six conditions: natural speech (panel A), UTVS (panel B), AMTVS25 (panel C), AMTVS50 (panel D), AMTVS100 (panel E), AMTVS200 (panel F), CAMTVS (panel G).

Figure 6. Mean phoneme accuracy across age groups as a function of modulation rate. Error bars represent 95% confidence intervals.

Figure 7. Mean phoneme accuracy scores for UTVS sentences as a function age. Error bars represent 95% confidence intervals.

Figure 8. Modulation-based intelligibility changes across age. Error bars represent 95% confidence intervals.

Figure 9. Modulation-based phonetic accuracy changes for the combined age groups. Error bars represent 95% confidence intervals.

APPENDICES

APPENDIX

TABLE OF CONTENTS

CHAPTER 1

Introduction

Speech perception in natural environments is a complicated task that requires processing at many different levels. Before word recognition can occur, the listener must be able to separate the speech of interest from background noise. In this case, noise is defined as any sound other than the one the listener wants to hear. Consider, for example, two people having dinner in a restaurant. The sounds reaching their ears may include each other's voice (voices they want to hear) and voices of diners at other tables and those of restaurant staff (voices they don't want to hear), as well as the clinking of dinnerware, the music playing in the background, and the sounds of chairs and shoes on the tile floor that adds just the right ambiance to the setting. Although on a spectrogram, the combined sounds would be indistinguishable from one another, in most instances humans are able to segregate all of these sounds from one another and focus on the sounds that interest them.

Cherry's 1953 research (c.f. Bregman, 1990) termed the difficulty of understanding one voice in the midst of many voices "the cocktail party problem". He reported a number of factors that affected how well subjects were able to segregate two voices that were presented simultaneously. One factor was whether the voices were separated in space (in his experiments they were sent to two different earphones). Another factor was predictability. If the two voices were sent to the same ear, Cherry reported that people were better able to follow a single voice if what was being said was predictable based on what preceded it. Cherry also proposed that factors such as voice quality, pitch, accent, and rate of speech could affect the ability to distinguish separate voices.

The ability to separate the auditory foreground from background has been referred to as "auditory object formation" (Moore, 1989). Many of the processes that contribute to the formation of auditory objects may also be responsible for our ability to group sounds into a coherent speech signal. Bregman (1990) states that an analysis of the auditory scene is the first step in the process of grouping items auditorally. Lass (1996) describes auditory scene analysis as "the process by which the brain reconstructs the external world through intelligent analysis of acoustic cues and information" (p. 392). This process includes both simultaneous and sequential grouping cues (Moore, 2003). Simultaneous cues allow a listener to group sounds that are occurring at the same time as coming from a single source. These include cues such as common fundamental frequencies, similarity of onsets and offsets, harmonicity, and comodulation. Sequential cues are those that allow a listener to group sounds that occur over time as belonging to a single source. These include cues such as continuity of pitch, interaural time differences, and temporal order. Factors specific to speech, such as phonetic and semantic relations or listener expectations, also provide information useful for grouping speech sounds into an auditory object (Miller et al., 1951). As Moore (2003) states, "the peripheral auditory system acts as a frequency analyzer, separating the different frequency components in a complex sound. Somewhere in the brain, the internal representations of these frequency components have to be assigned to their appropriate sources." (p. 274)

The ability to perceive speech as an auditory object, separate from its environment is thought to be a necessary state in the process of speech perception that contributes to higher level processes such as phonemic or syntactic interpretation. While auditory object formation has been studied in adults, to date, there have been no studies conducted describe the developmental time course of the

ability to use auditory object formation. Since children and adults differ on many auditory tasks, it is reasonable to question whether there may be developmental aspects to auditory object formation.

Studies have shown that children differ from adults on a number of nonspeech auditory tasks. For example, investigators have reported developmental changes in detection thresholds both in quiet (Schneider, Trehub, Morrongiello, & Thorpe, 1986; Trehub, Schneider, Morrongiello, & Thorpe) and in noise (Allen & Wightman, 1995; Buss, Hall, Grose, & Dev, 1999; Hall & Grose, 1991; Nozza & Wilson, 1984; Schneider, Trehub, Morrongiello, & Thorpe, 1989; Wightman, Callahan, Lutfi, Kistler, & Oh, 2003), including masking level difference (MLD) tasks (Hall & Grose, 1990; Grose, Hall & Dev, 1997; Hall, Buss, Grose, & Dev, 2004), as well as forward, backward, and simultaneous masking tasks (Buss, Hall, Grose, & Dev, 1999; Hartley, Wright, Hogan, & Moore, 2000; Hill, Hartley, Glasberg, Moore, & Moore, 2004). Developmental changes have been reported in temporal resolution (Allen, Wightman, Kistler, & Dolan, 1989; Trehub, Schneider, and Henderson, 1995), and in comodulation masking release (CMR) (Veloso, Hall, & Grose, 1990; Hall, Grose, & Dev, 1997), while frequency resolution/frequency selectivity appears adult-like at a very young age (Allen, Wightman, Kistler, & Dolen, 1989; Oh, Wightman,& Lufti, 2001; Veloso, Hall, & Grose, 1990).

Investigators also have shown differences between children and adults on a variety of speech perception tasks. Age differences have been shown in speech perception in noise and reverberation (Hall, Grose, Buss, & Dev, 2002; Johnson, 2000; Litovsky, 1997; Nozza, Rossman, Bond, & Miller, 1990; Neuman & Hochberg, 1983, Fallon, Trehub, & Schneider, 2000) and in the presence of reduced spectral cues (Eisenberg, Shannon, Martinez, Wygonski, & Boothroyd, 2000),

Numerous studies have shown that children differ from adults in their use of temporal and contextual cues in speech perception (Elliott, 1979; Elliott Busse, Partridge, Ruper, & DeGraffe, 1986; Elliott, 1986; Nittrouer and Boothroyd, 1990). In addition, children have been shown to require greater bandwidths and audibility to perceive some phonemes (Kortekaas & Stelmachowicz, 2000; Stelmachowicz, Pittman, Hoover, & Lewis, 2001). Investigators have shown that children and adults differ in the weights they assign to some acoustic parameters of speech and that these weights change as children gain more experience with their native language (Mayo, Scobbie, Hewlett, & Waters, 2003; Nittrouer, 1996; Nittrouer & Miller, 1997; Nittrouer & Crowther, 1998).

Theories regarding the underlying causes of reported developmental differences vary. Explanations include peripheral factors such as auditory sensitivity (Schneider et al., 1986, 1989; Trehub, Schneider, Morrongeillo, & Thorpe, 1988), as well as non-peripheral factors such as central auditory processes (Allen & Wightman, 1994, 1995) attention or memory (Allen & Wightman, 1994; Buss et al., 1999; Hill et al., 2004; Wightman et al., 2003), experience with language (Eisenberg et al., 2000), or poor information extraction (Grose et al., 1997). Fallon et al. (2000) point out that "the typical tasks used in speech-identification studies may pose disproportionate difficulty for young children" (p.3023).

Because speech is a complex signal that often contains numerous cues that can be used for perception (Denes, 1955; Repp, 1982), the listener may rely on only one or two cues in adverse listening situations (e.g., masking by background noise). Therefore, the robust nature of speech makes it difficult to evaluate the contribution of individual cues to perception. What is needed is an impoverished form of speech that does not contain many of the grouping cues known to aid speech perception.

Time-varying sinusoidal (TVS) speech (Remez, Rubin, Pisoni, & Carrell, 1981; Carrell & Opie, 1992; Barker & Cooke, 1999) consists of three to four constant amplitude, time-varying sinusoids that follow the center frequencies of formants of naturally spoken utterances. TVS speech does not contain the fundamental frequencies, harmonic structure, formant frequency transitions, or short-term spectral cues found in natural speech (Figure 1). As a result, acoustic cues thought to promote auditory object formation may be systematically investigated using TVS Speech.

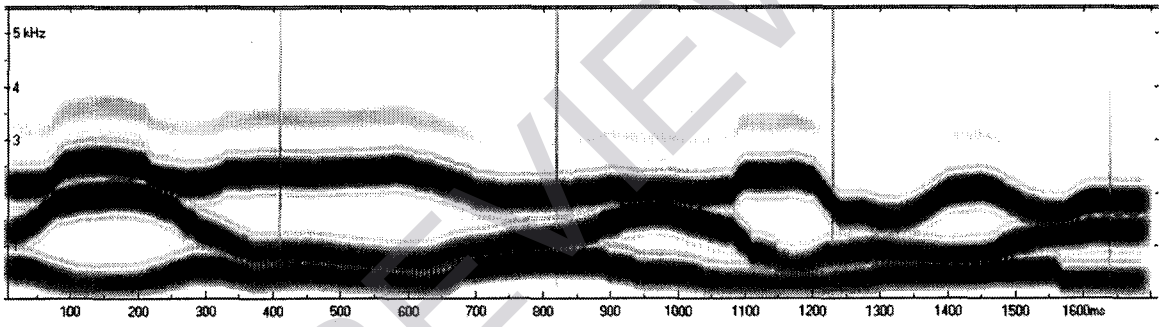

Figure 1. Narrowband spectrogram of the TVS sentence, "The yellow lion roared."

Although TVS speech has been used extensively in studies involving adults (e.g., Barker & Cooke, 1999; Carrell & Opie, 1992; Goh, Pisoni, Kirk, & Remez, 2001;Sheffert, Pisoni, Fellowes, & Remez, 2002; Remez, Rubin, Pisoni, & Carrell, 1981; Remez & Rubin, 1990; Rosner, Talcott, Witton, Hogg, Richardson, Hansen, & Stein; 2003), it's use with children is limited (Serniclaes, Sprenger-Charolles, Carré, & Demonet, 2001). Serniclaes et al. (2001) used TVS syllables to compare categorical perception for children with dyslexia and children who were normal readers (all were 13 years of age at the time of testing). The test syllables varied along a place-of-

articulation continuum. Two versions of the TVS signals were tested: unmodulated and the same signals amplitude modulated. Serniclaes et al. reasoned that the addition of amplitude modulation "had the effect of giving the sounds the equivalent of a voice pitch and made them immediately appear as speech-like sounds" (p. 389). Signals were presented in three conditions. In the first, subjects heard the unmodulated TVS syllables which were described as "electronic whistles". In the second, the same syllables were presented but the subjects were told that they were the syllables /ba/ or /da/ "pronounced in a special way by the computer, just as by Martians" (p. 399). In the third condition, the amplitude modulated TVS syllables were presented and the subjects were told that they would be "pronounced a bit as though a little French child was speaking" (p. 399). Subjects listened to pairs of syllables and their task was to indicate whether the stimuli were the same or different. Results revealed that children with dyslexia demonstrated better discrimination than the average readers, in that they showed better within category discrimination. As a result, they demonstrated less categorical perception. This was more noticeable for the speech-like than nonspeech condition. Therefore, although TVS stimuli have been used in one developmental study to date, it was limited to syllables rather than sentences. Moreover, this study examined perception across groups based on reading ability, not age.

The goal of the present investigation, was to determine how or if children between the ages of 4 and 13 years use amplitude modulation (AM) in speech perception, particularly in the formation of auditory objects. Because the study of auditory grouping in speech perception is complicated by the fact that speech contains many redundant acoustic characteristics, time-varying sinusoidal (TVS) speech was used. New stimuli were developed that would be age-appropriate for the

children in this study. Speech recognition of TVS sentences (unmodulated, amplitude comodulated at four frequencies, and amplitude modulated using conflicting frequencies) was measured. It was hypothesized that children would be less able than adults to use AM as an auditory grouping cue and that children's ability to use this information would improve with age.

CHAPTER II

Literature Review

To understand how listeners form auditory objects from the many sounds that reach their ears, one must begin by examining the cues that facilitate auditory grouping. For the present discussion, the focus will be on simultaneous grouping cues. That is, those cues that allow a listener to group sounds that are occurring at the same time as coming from a single source.

Auditory Grouping

Fundamental Frequency

Fundamental frequency may play a role in the grouping or separation of speech sounds such that those with the same or similar fundamental frequencies are heard as one sound (Assman and Summerfield, 1987; 1990; Bregman, 1990; Chalikia And Bregman, 1989; Darwin, 1981). In a series of experiments, Darwin (1981) examined the role of fundamental frequency on the perceptual grouping of speech sounds. In the first three experiments, listeners were able to identify three-formant vowels and two-formant diphthongs with the same accuracy regardless of whether the formants had the same or different fundamental frequencies. A fourth experiment used four widely spaced formants that could form two different three-formant syllables (/ru/ or /li/). Results revealed that when all four formants were on the same fundamental frequency, the majority of subjects heard /ru/. When the fundamental frequency of the second formant was different than the other three, the majority of subjects heard /li/, suggesting that this second formant was not grouped with the others. Darwin concluded that, although fundamental frequency can influence the grouping of speech sounds it is not "a necessary or sufficient condition for formants to be grouped into a common speech category" (p. 185).

Chalikia and Bregman (1989) investigated the effect of two auditory grouping cues on perception of vowels presented simultaneously: fundamental frequency and "common fate". Subjects heard simultaneously presented vowel pairs and were asked to identify vowels they heard under conditions of varying separation in fundamental frequency and changes in pitch contour across time (either steady state or gliding). Results revealed that increasing differences in fundamental frequency improved listeners' abilities to hear two separate vowels. In addition, gliding pitch contours, especially crossing glides (the pitch contour of one vowel rose and the other fell over time) improved performance.

Onset/Offset Asynchony

Synchronicity of onset and offset also has been identified as a potential cue to auditory grouping (Bregman, 1990; Darwin, 1981, 1984). In the study described above, Darwin (1981) also investigated the effect of common onset or offset on auditory grouping of speech sounds using synthesized three-formant vowels. Onset or offset were varied by either starting the lowest or highest formant 100 ms ahead of the other two formants or ending it 100 ms after. Darwin found that staggering the onset or offset of one of the formants resulted in subjects hearing two rather than one sound. This effect was more pronounced when the higher formants were staggered. However, staggering the onset or offset time of one formant relative to the others did not affect listeners' abilities to identify the vowels even though they perceived more than one sound.

Darwin (1984) further investigated the effect of onset and offset on the perception of vowels. A 500 Hz tone with equal amplitude and phase as the original harmonic within the vowel was added to 11 synthetic vowels across the continuum from /I/ to /E/. The tone was on for the duration of the vowel but, across conditions, it

began or ended relative to the vowel as follows: beginning at the same time, 32 ms or 240 ms before and ending at the same time; beginning at the same time and ending at the same time, 32 ms or 240 ms after. Subjects' task was to identify the vowel that was presented as either /I/ or /E/. If the added energy did not change perception (i.e. was grouped with the original vowel) there would be no change in the identification functions relative to the stimuli without the added tones. Results indicated a significant shift in the functions when the added energy started and ended at the same time as the original vowel. There were no significant changes in the function relative to the original stimuli when the added energy started 32 or 240 ms before the vowel. When the added energy ended 240 ms after the vowel, the function shifted from that obtained in the simultaneous condition toward that obtained in the original (no added energy) condition. Although there was a slight shift in the 32 ms offset condition, it was not significant. Overall, the results indicated that harmonics that start or stop before or after the vowel will be perceived as separate auditory objects. These findings suggest that small differences in time are important for the formation of auditory objects.

Comodulation Masking Release

A tone that is presented at subthreshold levels and is centered in an amplitude modulated narrowband noise may become audible if another band of noise is added at a different frequency. This occurs only if the second band of noise is amplitude modulated at the same rate and phase as the first band of noise. One interpretation of this phenomenon, known as comodulation masking release or CMR (Hall, Haggard, & Fernandes, 1984), is that the noise bands are grouped together by their common amplitude modulation, making the excluded tone more audible.

Hall and Grose (1990) investigated the relation between CMR and auditory grouping. Earlier studies had shown that CMR could be reduced if additional noise bands that were not comodulated with the on-signal band were included along with the comodulated off-signal bands. Hall and Grose found that the negative effect of these "deviant" bands was be reduced as more were added and/or when there was onset and offset asynchrony with the on-signal and comodulated bands of noise. They concluded that their "results suggest that CMR analysis may occur after spectral components have been organized perceptually (e.g., by common amplitude modulation and/or onset/offset asynchrony)" (p. 125).

Grose and Hall (1992) examined comodulation masking release for speech stimuli using band-pass filtered vowels, consonants, and sentences. Adult listeners were trained to recognize the filtered speech without maskers prior to testing. Results for vowels revealed an improvement in recognition threshold in the presence of additional masking bands. However, no improvement in threshold was seen for the consonants. When testing using sentence material, CMR was measured both for detection and for recognition tasks. Results revealed an average CMR of 4.6 dB for detection of sentences. Conversely, no CMR was noted for the recognition task. The authors concluded that their results supported CMR for speech detection (threshold) tasks but not for speech recognition (suprathreshold) tasks.

In contrast, Kwon (2002) reported a comodulation masking release effect in a consonant recognition task. In a preliminary experiment, Kwon found a CMR effect for listeners identifying manner features of consonants but not place features. Further testing with a larger set of consonants revealed a small but statistically significant effect of CMR in consonant recognition. However, in the latter experiment no difference between place and manner features was noted.

Time Varying Sinusoidal Speech

Recall that time-varying sinusoidal (TVS) speech consists of three to four constant amplitude, time-varying sinusoids that follow the center frequencies of formants of naturally spoken utterances. It does not contain the fundamental frequencies, harmonic structure, formant frequency transitions, or short-term spectral cues found in natural speech. Remez et al. (1981) evaluated independent groups of listeners' perception of a three-tone TVS sentence. One group was told they were listening to computer-generated speech and another group was told to report their spontaneous impression of the stimuli without being told that it was speech. Results revealed that the majority of listeners who were not told they were listening to speech did not hear speech. However, those who were told they would be listening to speech were much more likely to hear and understand at least parts of the three-tone sentence. The authors concluded that listeners were able to understand the sinusoidal signals "because the linguistic information, although not carried by acoustic elements producible by a vocal tract, is preserved in the time-varying relational structure of the stimulus pattern" (p. 949). It has been argued that sinusoidal signals also carry "Gestalt-style cues to integration" in that the tones depict the common onsets and offsets and continuity of the original speech signal (Darwin, n.d.; Ellis, 1996).

Carrell and Opie (1992) used TVS speech to determine the contribution of amplitude comodulation to auditory grouping in speech perception. They compared recognition scores for unmodulated TVS (UTVS) sentences to sentences where the three sinusoids were amplitude modulated simultaneously (AMTVS) at 100 Hz. Results revealed improved intelligibility for the AMTVS sentences, supporting the hypothesis that the modulation served as a mechanism for grouping the three sinusoids as one auditory object, thereby increasing intelligibility.

Figure 2 illustrates narrowband spectrograms of natural speech (upper panel), UTVS speech (middle panel), and AMTVS speech (lower panel) for the utterance "The swan dive was far short of perfect." The narrow horizontal lines in the natural speech represent harmonic structure. The only energy present in the UTVS sentence is at the center frequencies of the first three formants. The AMTVS speech more closely resembles the natural speech because of the 100-Hz sidebands flanking each of the center frequencies. It is important to note, however, that these sidebands are not harmonically related to a fundamental frequency as is the case in the natural speech.

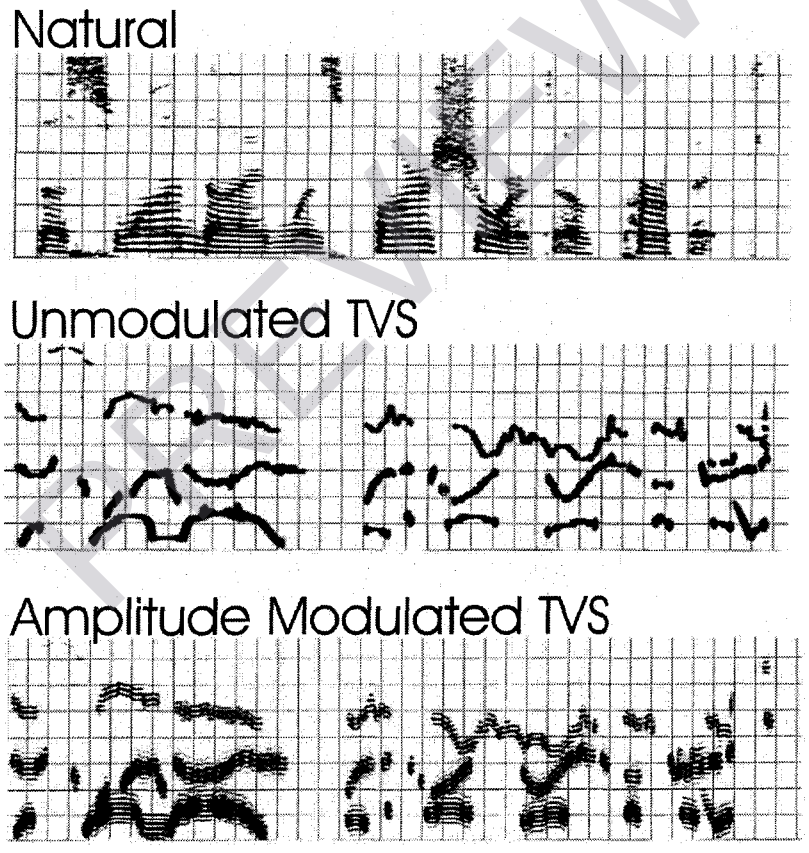

Figure 2. Narrowband spectrograms of the sentence "The swan dive was far short of perfect" representing natural speech, unmodulated TVS speech, and TVS speech comodulated at 100 Hz.