Development, Perceptual Evaluation, and Acoustic Analysis of Amplitude-based F0 control in Electrolarynx Speech

by

Yoko Saikachi

B.A. French Literature, University of Toyo, 1999

M.S. Health Science, University of Tokyo, 2002

MSc Speech and Language Processing, University of Edinburgh, 2003

Submitted to the Harvard-MIT Division of Health Science and Technology in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY
IN SPEECH AND HEARING BIOSCIENCE AND TECHNOLOGY
AT THE
MASSACHUSETTS INSTITUE OF TECHNOLOGY

SEPTEMBER 2009

Signature of Author: _____
Harvard-MIT Division of Health Sciences and Technology
August, 20, 2009

Certified by: _____
Robert E. Hillman, Ph.D., CCC-SLP
Associate Professor of Surgery, Harvard Medical School
Thesis co-supervisor

Certified by: _____
Kenneth N. Stevens
Clarence J. LeBel Professor of Electrical Engineering and Computer Science
Professor of Health Sciences and Technology
Thesis co-supervisor

Accepted by: _____
Ram Sasisekharan, PhD
Edwin Hood Taplin Professor of Health Sciences & Technology and Biological Engineering
Director, Harvard-MIT Division of Health Sciences and Technology

Development, Perceptual Evaluation, and Acoustic Analysis of Amplitude-based F0 control in Electrolarynx Speech

By
Yoko Saikachi

Submitted to the Harvard-MIT Division of Health Sciences and Technology Speech and Hearing Biosciences and Technology Program on August, 2009 in Partial Fulfillment of the Requirements of the Degree of Doctor of Philosophy in Speech and Hearing Bioscience and Technology

ABSTRACT

An Electrolarynx (EL) is a battery-powered device that produces a sound that can be used to acoustically excite the vocal tract as a substitute for laryngeal voice production. ELs provide laryngectomy patients with the basic capability to communicate, but current EL devices produce a mechanical speech quality which has been largely attributed to the lack of natural fundamental frequency (F0) variation. In order to improve the quality of EL speech, the present study aimed to develop and evaluate an automatic F0 control scheme, in which F0 was modulated based on variations in the root-mean-squared (RMS) amplitude of the EL speech signal. Recordings of declarative sentences produced by two male subjects before and after total laryngectomy were used to develop procedures for calculating F0 contours for EL speech, and perceptual experiments and acoustic analyses were conducted to examine the impact of F0 modulation on the quality and prosodic function of the EL speech. The results of perceptual experiments showed that modulating the F0 of EL speech using a linear relationship between amplitude and frequency made it significantly more natural sounding than EL speech with constant F0, but also revealed some limitations in terms of communicating linguistic contrasts (distinction between question vs. statement and location of contrastive stress). Results are interpreted in relation to the acoustic characteristics of F0 modified EL speech and discussed in terms of their clinical implications and suggestion for improved algorithms of F0 control in EL speech.

Thesis Co-Supervisor: Robert Hillman, Ph.D., CCC-SLP
Title: Co-Director and Research Director, Center for Laryngeal Surgery and Voice Rehabilitation, MGH; Associate Professor of Surgery, HMS; Professor of Communication Sciences and Disorders, Massachusetts General Hospital Institute for Health Professions; HST Faculty

Thesis Co-Supervisor: Kenneth N. Stevens
Title: Clarence J. LeBel Professor of Electrical Engineering, Emeritus, Department of Electrical Engineering and Computer Science, MIT; Professor of Health Sciences and Technology, Emeritus, MIT

ACKNOWLEDGEMENTS

I would like to thank so many people for their support during my six years at MIT, but first and foremost, I would like to thank my thesis co-supervisors, Professor Robert Hillman and Professor Kenneth Stevens for their patient and continued encouragement, assistance, and invaluable advice and support. It was really fortunate to work as their students during my entire time for Ph.D. and I will always remember this time period with great appreciation and a lot of meaning. I would also like to thank my thesis committee chair, Professor Adam Albright, whose perceptive and encouraging comments were precious in achieving the goal of this study.

I am also thankful to many people whom I met in the course of the research. In particular, I am grateful to Geoff Meltzner and Harold Cheyne for being always available for technical issues and spending a lot of time and energy reading though the draft of the grant application and journal paper. I also want to thank the current and former members of the speech communication group at MIT and members of Center for Laryngeal Surgery and Voice Rehabilitation at MGH, especially, Jennifer Bourque, Anatoly Goldstein, Marie Jett, James Heaton, Jim Kobler, Yoshihiko Kumai, Arlene Wint, Stefanie Shattuck-Fufnagel, Seth Hall, Satra Ghosh, Hiroya Sadao, Janet Slifka, Helen Hanson, Takayuki Arai, and Joe Perkell for all of their assistance in recruiting participants for perceptual experiments and learning the hardware and software used in the lab. I am also deeply grateful to the Japan-US Educational Commission and Mizuho Iwata for their support.

I would also like to give special thanks to graduate students in speech and hearing program and graduate members of the Research Laboratory of Electronic's Speech Communication Group and Center for Laryngeal Surgery and Voice Rehabilitation at MGH, especially Xuemin Chi, Elisabeth Hone, Xiaomin Mou, Julie Yoo, Sherry Zhao, Nancy Chen, Tony Okobi, Steven, Lulich, Cara Stepp, Daryush Mehta, Asako Masaki, Brad Buran, and Prakash Srinivasamurthy Ravi. I am also indebted many friends who have always been supportive throughout my life at MIT and in my projects, particularly Koichi, Rong, Yu, Tor, Shioulin, Ai, Edward, Ji-Eun, Manshi, Natsuko, Matt, and Yuki. And lastly, I owe many, many thanks for the love and support that my family and friends in Japan gave me over the years.

TABLE OF CONTENTS

List of Figures

.

List of Tables

Chapter 1

Introduction

1.1. Laryngectomy and Alaryngeal Speech

Laryngectomy, usually necessitated by laryngeal cancer, involves the surgical removal of the entire larynx. This also causes the respiratory tract to be separated from the vocal tract, including the oral and nasal cavities. Consequently, breathing now occurs via the tracheostoma, an opening that is created by attaching the trachea to the skin in the neck. Due to the loss of larynx and the vocal folds, laryngectomy patients are no longer be able to produce normal voice. However, the vocal tract and articulators are still intact and it is possible to replace phonation by an alternative voicing source in order to speak.

The speech produced by the alternative voicing source is called alaryngeal speech and there are three main types; esophageal, tracheo-esophageal (TE), and Electrolarynx (EL). The focus of this dissertation is on enhancement of speech produced by EL. This chapter will therefore introduce basic background related to main acoustic deficits associated with EL speech, followed by a summary of the previous studies on the enhancement of the quality of EL speech. Lastly, the general objectives and overview of the dissertation will be presented.

1.2. Electrolarynx Speech

1.2.1. Electrolarynx

EL is a battery-powered device that produces a sound that can be used to acoustically excite the vocal tract as a substitute for laryngeal voice production. In the United States, the prevalence of EL use among patients is as high as 85% at one-month post-laryngectomy (Hillman, Walsh, Wolf, Fisher, & Hong, 1998), with multiple studies reporting longer term use of an EL as a primary mode of communication by more than half of laryngectomy patients (Gray & Konrad, 1976; Hillman et al., 1998; Morris, Smith, Van Demark, & Maves, 1992). Two ELs are currently available for use by laryngectomy patients: the neck-type (transcervical or trancutaneous) and mouth-type (transoral or intraoral). The current study focused on a neck-type device because this is by far the most commonly used type of EL.

1.2.2. Acoustic deficits of EL speech

ELs provide laryngectomy patients with the basic capability to communicate verbally (using oral speech production), provided conditions are sufficiently favorable (e.g., there is minimal competing noise, the listener has normal hearing and is familiar with EL speech). However, EL speech contains persistent acoustic deficits that result in reduced intelligibility and contribute to its "mechanical" or "non-human" (robotic) speech quality that often draws undesirable attention to the user. EL users have a particularly difficult time communicating with individuals who are unfamiliar with EL speech, which can make telephone use especially problematic. The main acoustic deficits associated with EL speech are (a) lack of normal fundamental frequency (F0) variation (Ma, Espy-Wilson, & MacAuslan, 1999; Meltzner & Hillman, 2005; Uemi, Ifukube,

Takahashi, & Matsushima, 1994); (b) the presence of the directly radiated signal (i.e., the buzz from the EL that is not filtered by the user's vocal tract but radiates directly to the listener) (Cole, Stridharan, Moody, & Geva, 1997; Espy-Wilson, Chari, MacAuslan, Huang, & Walsh, 1998; Liu, Zhao, Wan, & Wang, 2006; Niu, Wan, Wang, & Liu, 2003; Pandey, Bhandarkar, Bachher, & Lehana, 2002; Pratapwar, Pandey, & Lehana, 2003): and (c) an improper source spectrum (Qi & Weinberg, 1991; Weiss, Yeni-Komshian, & Heinz, 1979).

The contributions of precise F0 control in natural speech to prosodic and segmental contexts (i.e., signaling stress, syntactic and emotional information) are not possible with current EL technology. External placement of the EL and its interface with the neck contribute a direct-path signal to the intended speech signal that is transmitted via the vocal tract. The last deficit, an improper source spectrum, arises from a combination of the EL transducer design and the properties of the neck tissue. Existing EL transducers use a piston hitting a plastic disk to produce a mechanical "buzz". For example, the waveform generated by the Servox EL consists of a train of sharp impulses followed by highly damped oscillations (Qi & Weinberg, 1991). The spectrum of this signal is characterized by a broadband output with a spectral maximum around 2 kHz, and a substantial deficit in energy below 400-500Hz. This impulse-train-like signal then passes through the neck tissue transfer function, characterized by Meltzner et al. (2003) as low-pass, with a constant maximum gain between 100 Hz and a corner frequency in the range 200-400 Hz, then rolling off with a slope about -9 dB/octave until 3000 Hz where the magnitude becomes constant until 4000 Hz. Meltzner et al. (2003) also showed that other important properties may contribute to the poor quality of EL speech both in

terms of the improper acoustic characteristics and location of the voicing source provided by the EL transducers, and modifications in vocal tract transfer functions due to the impact of the laryngectomy operation on the upper airway.

1.2.3. Importance of F0 in EL speech

Several studies have demonstrated that significant improvements in EL speech could be accomplished by adding appropriate control of F0. Some of the work has illustrated the linguistic deficits caused by a lack of F0 control (Gandour & Weinberg, 1983, 1984; Weinberg & Gandour, 1986). For example, Gandour and Weinberg (1983) conducted perceptual experiments in order to determine the degree to which EL speakers were able to achieve intonational contrasts. Results showed that users of the electrolarynx were generally unable to achieve intonational distinctions with a flat F0 contour, indicating the critical role of F0 modulation. It has been also indicated that F0 cues may be useful for signaling segmental information. Recent work has shown that speech intelligibility was significantly better with variable intonation produced by a laryngectomy patient using an EL with manual F0 control (Tru-Tone, Griffin laboratories) than it was with either the resynthesized flat intonation or the fixed-frequency intonation (Watson & Schlauch, 2009).

Lack of adequate F0 control has been shown to be even more detrimental to the intelligibility of EL users who speak tone languages such as Thai, Mandarin, and Cantonese (Gandour, Weinberg, Petty, & Dardarananda, 1988; Liu, Wan, Wang, & Lu, 2006; Ng, Gilbert, & Lerman, 2001), where F0 contours contributed most to the perception of meaning among the three main acoustic cues (F0 contour, duration, and

amplitude contour) (Ng et al., 2001). More recent work has examined the impact of aberrant acoustic properties on the quality of EL speech. Meltzner and Hillman (2005) demonstrated that the addition of normal F0 variation was associated with the largest improvements in the "naturalness" of EL speech, as compared to other acoustic enhancements (compensation for low frequency deficit and reduction of noise radiated directly from the device). Ma et al. (1999) developed a post-processing scheme in which a cepstral-based method was used to replace the original F0 contour of EL speech with a normal F0 pattern and showed that adding F0 variation clearly improved naturalness of EL speech. Although this post-processing technique was promising, its practical applications are limited because it requires pre-recording EL speech and cannot be implemented in real-time.

1.2.4. Previous work on controlling F0 in EL speech

Adding the proper F0 variation to EL speech in real-time is very challenging because it would require the means to estimate what pitch the speaker intends to use (i.e., access to underlying linguistic and/or neural processes), or utilization of alternative signals or control sources (e.g., Kakita & Hirama, 1989; Sekey & Hanson, 1982; Uemi et al., 1994). In one such approach, Uemi et al. (1994) used air pressure measurements obtained from a resistive component placed over the stoma to control the fundamental frequency of an EL, but only 2 of 16 subjects were able to master control of the device. Other work has demonstrated the potential feasibility of accessing laryngeal neuromotor signals post-laryngectomy to use in controlling the onset, offset, and F0 of an EL. However, this general approach requires further testing and development, and may not be

effective in all EL users (Goldstein, 2003; Goldstein, Heaton, Kobler, Stanley, & Hillman, 2004; Heaton et al., 2004).

Other possibilities for controlling F0 in EL speech include implementing a fixed F0 contour (van Geel, 1982; Secom MYVOICE). For example, MYVOICE (Secom) produces a high F0 frequency when the switch of the EL is turned on, and the F0 then falls. This is more natural than a flat F0, but as the F0 variation pattern is fixed, the device cannot be used with the intention of adding various intonation patterns, and it may lead to confusion of the speaker's intent (e.g., a question with declarative prosody). There have been attempts to include manual control of F0 in the design of some EL devices (Choi, Park, Lee, & Kim, 2001; Galyas, Branderud, & McAllister, 1982; Kikuchi & Kasuya, 2004; Takahashi, Nakao, Kikuchi, & Kaga, 2005; Tru-Tone, Griffin Laboratories), but there is considerable skepticism that manual control (e.g., pushing a button with a finger) can successfully approximate the very precise and rapid adjustments in F0 that occur during normal speech production. Furthermore, learning to effectively control F0 manually may be particularly difficult for the majority of laryngectomy patients due to their advanced age.

1.3. Current Study

1.3.1. Overall goals of the dissertation

This dissertation describes one approach that we have been developing to automatically control the F0 of EL speech. We are proposing to modulate the F0 of EL utterances based on variation in the root-mean-squared (RMS) amplitude of the EL

speech signal. In previous acoustic studies of the speech of patients before (laryngeal speech) and after (EL speech) total laryngectomy, we found significant fluctuations in the amplitude of EL speech (Saikachi, Hillman, & Stevens, 2005). In particular, there was a gradual decrease of amplitude during vowels at the end of declarative utterances, which was similar to what we observed in the corresponding pre-laryngectomy speech. Furthermore, there were generally positive correlations between F0 and amplitude in pre-laryngectomy (laryngeal) speech (i.e., the shape of the amplitude-time curves were similar to the F0-time curves). Based on these observations and previous finding, we hypothesized that the amplitude variations in EL speech could be used as a basis for effectively predicting, and ultimately controlling, the F0 of EL speech in close to real-time.

This work is motivated by a long-term goal to develop a real-time speech processing technology to remedy the acoustic deficits in EL speech and thereby improve EL communication, and the quality of life, for laryngectomy patients. Enabling users to automatically control F0 may improve communication efficiency, improve social interaction, and enhance overall communication satisfaction. Furthermore, the findings obtained from this study may provide the basis for developing intervention strategies that improve prosodic control capabilities for communicative function. The overall goal of this investigation was to evaluate the viability of the proposed approach by: 1) developing procedures for estimating F0 based on the amplitude variations in EL speech, 2) evaluating the impact of amplitude-based modulation of F0 on the quality of EL speech in perceptual experiments, 3) determining the limitations of the developed

approach in communicating linguistic contrast, and 4) determining the acoustical basis of the perceptual characteristics of F0 modified EL speech.

1.3.2. Outlines of the dissertation

To achieve these general research goals, the experiments reported on in the different chapters concentrate on the development and evaluation of F0 control scheme. The dissertation is organized as follows:

The first part of Chapter two describes the motivations and procedures for developing the amplitude based F0 control scheme. The second part evaluates the developed control scheme in terms of improvement in overall naturalness using formal perceptual experiments.

Building on the results reported in Chapter two, Chapter three and four will more fully examine the ability and limitation of amplitude based F0 control in prosodic aspects. More specifically, Chapter three will examine the impact of F0 modulation on communicating linguistic contrast. Chapter four will examine the acoustic basis of the perceptual results obtained in Chapter three and the relationship between perceptual results and acoustic analysis will be discussed.

Chapter five will summarize overall results and discuss clinical implication and limitation of the current study followed by future research perspectives.

Chapter 2

Development and Perceptual Evaluation of F0 Control in EL Speech

2.1. Methods

2.1.1. Speech Recordings

In the present study, two declarative sentences from the "Zoo passage" produced by two male subjects (referred to as "speakers 1 and 2" hereafter) before and after total laryngectomy (pre-laryngectomy speech vs. EL speech) were selected from the recordings made for the Veterans Administration Cooperative Study # 268 (VA-CSP 268). Recording of subjects from this data set who had acceptable pre-laryngectomy voice quality have been particularly useful for assessing the acoustic differences between normal (laryngeal) and EL speech, and for providing acoustic "targets" to improve EL speech (Goldstein et al., 2004; Heaton et al, 2004; Meltzner, 2003; Meltzner & Hillman, 2005). Sentence 1 was "His sister Mary and his brother George went along, too." And sentence 2 was "You can see that they didn't have far to go." These declarative sentences were chosen because each one terminated with vowels in which amplitude decreased

consistently in both the pre-laryngectomy and EL speech of the two speakers (Saikachi et al., 2005).

The two speakers were chosen because they used EL speech as their primary mode of communication, the level of interference due to directly radiated EL noise was relatively low in their post-laryngectomy recordings, and their pre-laryngectomy speech was found to have relatively normal voice quality (tumor location minimally affected voice production). The two speakers both used a neck-placed Servox EL, but were recorded at different VA hospitals. Of the several post-laryngectomy recordings that were made for each speaker, only the final EL speech recordings were used in this study (30 months post-laryngectomy for speaker 1 and 12 months post-laryngectomy for speaker 2). All recordings were made in a quiet environment using a Marantz model 220 recorder and a Radio Shack model 33-1071 microphone, situated 6 to 12 inches from the speakers (Hillman et al., 1998). An audio signal acquisition and editing software package (Syntrillium Software's Cool Edit 2000) was used to digitize the speech at 32 kHz. For this study, the speech was appropriately low pass filtered and downsampled to 10 kHz.

2.1.2. Amplitude based F0 estimation

Figures 2.1 and Figure 2.2 show representative data from pre-laryngectomy and EL speech respectively, including the audio waveform, F0 contour, and RMS amplitude as a function of time during sentence 1. F0 was estimated using autocorrelation analysis (Markel & Gray, 1976). Both F0 and RMS amplitude were calculated every 5 ms over 40 ms intervals. Note that there is a fluctuation in amplitude over the whole utterance in both the pre-laryngectomy and EL speech.

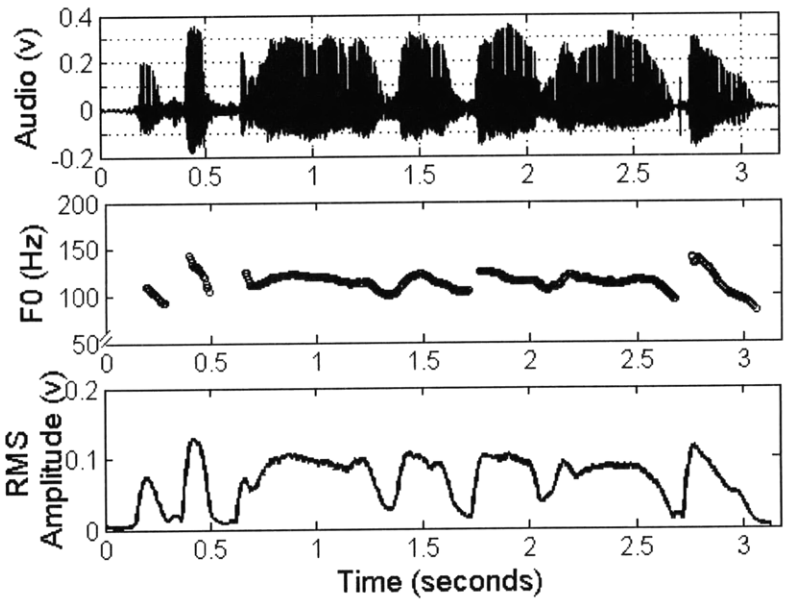


Figure 2.1: Audio waveforms, F0, and RMS amplitude over time for sentence 1, "His sister Mary and his brother George went along, too" recorded before laryngectomy (prelaryngectomy laryngeal speech).

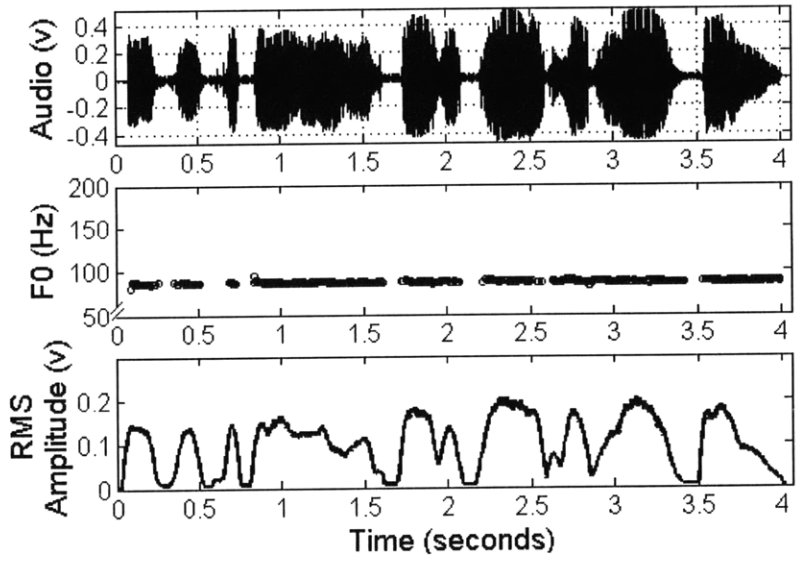


Figure 2.2: Audio waveforms, F0, and RMS amplitude over time for sentence 1, "His sister Mary and his brother George went along, too" recorded after laryngectomy (using an EL) by speaker 1.

The relationship between F0 and RMS amplitude in pre-laryngectomy speech served as the basis for using the amplitude variation of EL speech to generate an F0 contour. More specifically, for each sentence and each speaker, the linear regression coefficients (intercept and slope) between F0 and amplitude were calculated for the pre-laryngectomy sentences in order to model F0 as a function of RMS amplitude. Only the voiced parts in the sentences were included for the computation. F0 values that were miscalculated by the autocorrelation methods (either halved or doubled) were also excluded from the analysis.

Figure 2.3 shows F0 plotted against RMS amplitude for a pre-laryngectomy recording of speaker 1 producing sentence 1. Also shown in Figure 2.3 is the straight line that best fits the data, which clearly reflects the positive relationship between RMS amplitude and F0. Table 2.1 summarizes the regression coefficients and Pearson *r* correlation coefficients for both sentences produced by each of the two speakers. F0 and RMS amplitude were significantly correlated in each sentence ($p<.001$), and the regression coefficients varied depending on the speakers and sentences.

F0 contours for the EL speech were then derived from the RMS amplitude variation in EL speech using the following equation for each sentence and speaker:

$$\text{Estimated\_F0} = k_1 + k_2 \times \text{RMS\_amplitude} \qquad (1)$$

where k_1 and k_2 are respectively the intercept and slope of the regression coefficients obtained from analyzing the pre-laryngectomy speech. Figure 2.4 shows an example of an amplitude-based estimate of an F0 contour superimposed on the original F0 contour for sentence 1 produced by speaker 1 using an EL.

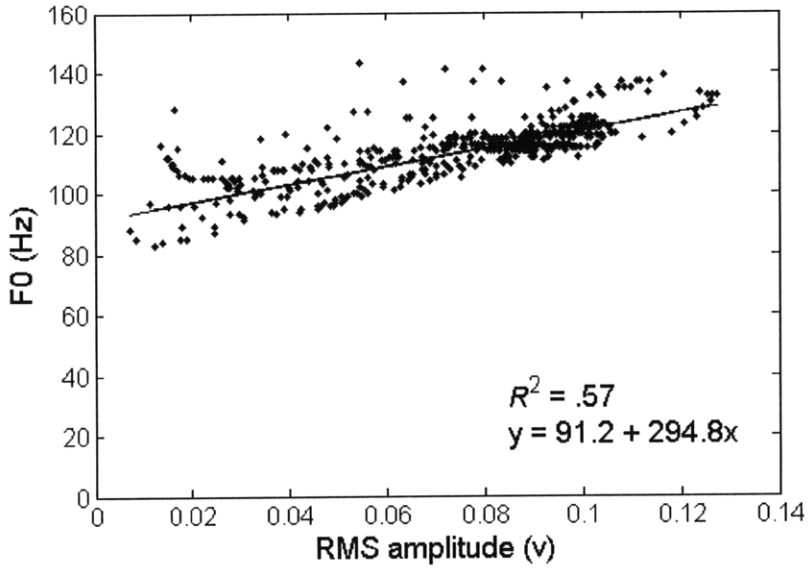


Figure 2.3: F0 versus RMS amplitude and linear regression for sentence 1 produced by speaker 1. Correlation coefficients and regression coefficients are shown at the bottom.

	Subject 1		Subject 2	
	Sentence 1	Sentence 2	Sentence 1	Sentence 2
Intercept (Hz)	91.2	102.6	92.0	91.4
Slope (Hz/Volts)	294.8	262.4	190.5	182.2
Correlation coefficients: R^2	.57*	.44*	.39*	.38*

Table 2.1: Values of intercept, slopes, and correlation coefficients for the different speakers and sentences *(*$p < .001$).

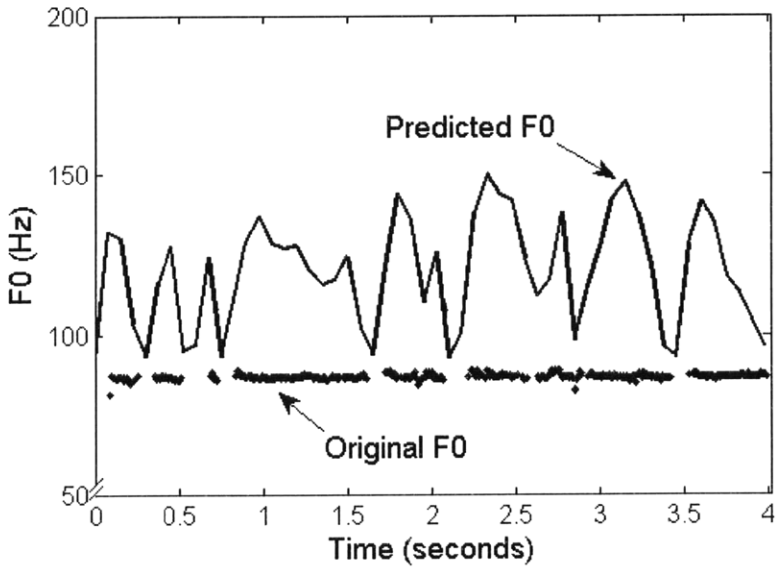


Figure 2.4: Measured original F0 and amplitude-based estimates of F0 as a function of time for sentence 1 produced by speaker 1.

2.2. Perceptual Evaluation

A perceptual experiment was conducted in order to determine whether the proposed approach for controlling F0 based on amplitude could significantly improve the naturalness of EL speech, and whether this approach was comparable to synthesizing EL speech with an F0 contour based on pre-laryngectomy speech.

2.2.1. *Generation of speech stimuli*

The first step in generating stimuli (speech tokens) to perceptually evaluate the impact of amplitude-based F0 modulation on the quality of EL speech was to synthesize EL speech using the Klatt formant synthesizer (KLSYN). KLSYN is a well established formant synthesizer that allows for direct control of both source and filter characteristics,

and it has been shown to have the capability of producing high quality copy synthesis for normal speech (Hanson, 1995; Klatt, 1980; Klatt, Chapter 3; Klatt & Klatt, 1990) as well as for pathological voices (Bangayan, Christopher, Alwan, Kreiman, & Gerratt, 1997). The motivation behind using this method is that synthesis can provide a tool through which the characteristics of EL speech and pre-laryngectomy speech can be compared at the level of the synthesis parameters, i.e., analysis-by-synthesis. After being parameterized, EL speech can be modified via individual or combinations of parameters to examine the resulting quality of the modified EL speech. The procedures for synthesizing EL speech are included in the Appendix. Once copy-synthesis of the original EL speech samples was accomplished, the F0 synthesis parameter was manipulated to produce EL stimuli with the desired F0 contours.

The overall scheme for generating speech tokens is shown in Figure 2.5. For each sentence-speaker condition, three versions of each sentence were generated from the copy-synthesized EL speech by simply modifying the F0 synthesis parameters:

(a) EL speech with constant F0 (EL_S);

(b) EL speech with F0 modulation based on the F0 contour of pre-laryngectomy speech (EL_f0n);

(c) EL speech with F0 modulation based on the amplitude of the EL speech (EL_f0a).

This resulted in 6 sentences per speaker, or a total of 12 sentences. The constant F0 values for the EL_S sentences were set to the average F0 of the pre-laryngectomy versions of the sentences, to minimize any confounding factor that could be related to differences in average F0 when comparing different stimuli. For the EL_f0a sentences,

the F0 was derived from the linear relationship between F0 and amplitude in the pre-laryngectomy speech samples as described previously using Eq. (1). The computed F0 was normalized such that the mean and variance of the F0 were matched to those in the pre-laryngectomy versions of the sentences.

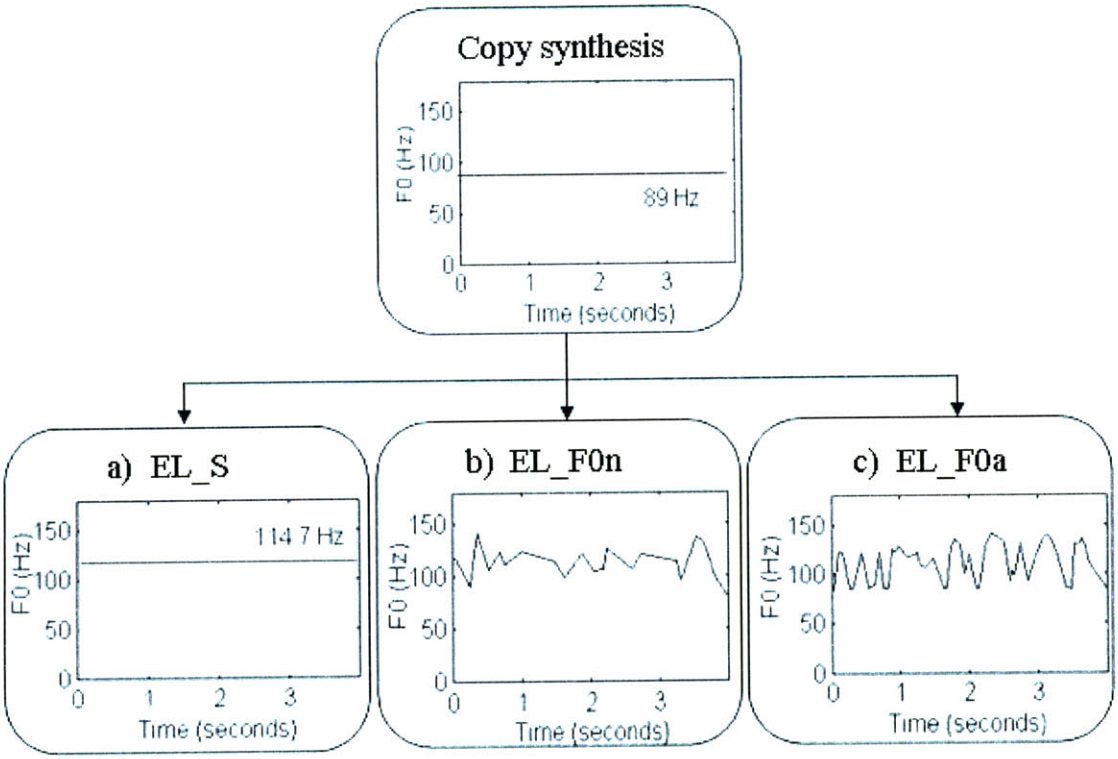


Figure 2.5: F0 synthesis contours for sentence 1 and speaker 1 that were used to generate the EL speech stimuli for the perceptual experiments. "EL_S" corresponds to copy-synthesized EL speech with constant F0. "EL_f0n" and "EL"f0a" are the EL speech with F0 modulations based on the pre-laryngectomy F0 contour and EL speech amplitude respectively.

For the EL_f0n sentences, adding the pre-laryngectomy F0 contours involved two steps. First, the pre-laryngectomy and EL sentences were time aligned using the Pitch-Synchronous Overlap-Add (PSOLA) algorithm (Moulines & Charpentier, 1990), such that the phones of both sentences had the same onset times and duration. The F0 contours

obtained from the time-scaled pre-laryngectomy versions of the sentences were then used to set the F0 synthesis parameters to generate the EL_f0n versions of the sentences.

2.2.2. Listeners

A group of 12 normal hearing graduate students recruited from MIT and the MGH Institute of Health Professions (six females and six males) served as listeners.

2.2.2. Experimental procedures.

The synthesized stimuli were perceptually evaluated using a combination of two approaches: the Method of Paired Comparisons (PC) (Meltzner & Hillman, 2005; Torgerson, 1957) and visual analog scaling (VAS).

Perceptual judgments were carried out within each of the four speaker-sentence conditions (2 speakers x 2 sentences = 4 conditions). Within each speaker-sentence condition, all combinations of pairs of the 3 synthesized speech tokens (3 pairs) were presented twice to listeners to total 6 paired-comparisons per condition. Thus, a total of 24 pairs of speech stimuli were presented to each listener (3 pairs x 2 repetitions x 4 conditions = 24), which resulted in a total of 288 listener responses for the entire study (24 stimulus pairs x 12 listeners = 288).

Before judgments were made within each of the four speaker-sentence conditions, all three speech tokens (EL_S, EL_f0a, EL_f0n) for that condition were played to the listeners to familiarize them with the quality of the different stimuli. The pre-laryngectomy speech sample associated with the condition being evaluated was also played as a reference for the perceptual judgments. This allowed the pre-laryngectomy

version of each sentence to act as an anchor so that all listeners would have a common frame of reference to make their judgments.

Each listener was seated in a sound-isolated booth and was instructed to indicate on a computer screen which of the two tokens in each pair sounded most like normal, natural speech. Then the listener was asked to rate how different the chosen token was from normal natural speech using a VAS on the computer screen. The VAS was 10 cm long, with the left end labeled *Not At All Different* and the right end labeled *Very Different*. The presentation order of four speaker-sentence conditions was randomized for each listener. Participants were allowed to listen to the pre-laryngectomy speech token associated with each condition (anchor) as often as they wanted during both PC and VAS components of the experiment.

2.2.4. Data analysis

The PC data were first analyzed by conducting binomial testing in order to test the significance of the results. The PC data were also converted to scale rankings using Thurstone's Law of Comparative Judgment (Thurstone, 1927), in which speech tokens that were most consistently judged to sound more like normal, natural speech across all listeners were given a higher scale value (Meltzner & Hillman, 2005). The data from the VAS procedure were analyzed by computing the distance in centimeters from the left end of the VAS. These distances were used to calculate a mean distance for each speech type and taken as an estimate of how different a listener judged a speech token to be from natural, normal speech.

2.2. Results

The reliability of listeners was evaluated by calculating the percentages of agreement in preference judgments made by each listener in response to the repeated presentation of all token pairs. The average intralistener agreement across all four speaker-sentence conditions (speaker1-sentence1, speaker1-sentence2, speaker2-sentence1, & speaker2-sentence2) for the PC task was 80.0 ± 16.1% (the range was 50-100%), using an exact agreement statistic (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). Intralistener agreement across all four conditions for the VAS task was evaluated using Pearson's *r* and was .83 ± .16 (the range was .52-.99).

The PC response data are summarized in Table 2.2. Shown are the total number and percentage of times that listeners judged each of the three different speech tokens to sound more normal or natural than the other two tokens in paired comparisons. The binomial test showed that there was a significant overall preference by the listeners for the F0 modulated EL speech (EL_f0a and EL_f0n tokens) as compared with the EL speech having constant F0 (EL_S tokens) ($p<0.01$). The exception was the EL_f0a vs. EL_S token pair for sentence 2 produced by Speaker 1. Conversely, there was no significant preference for the EL_f0n tokens over EL_f0a tokens.

Paired Comparison	Subject 1		Subject 2		Overall
	s1	s2	s1	s2	
EL_f0a vs. EL_S	95.8% 23/24*	70.8% 17/24	95.8% 23/24*	95.8% 23/24*	89.6% 86/96*
EL_f0n vs. EL_S	88.0% 21/24*	91.0% 20/22*	92.0% 22/24*	100.0% 23/23*	96.0% 89/93*
EL_f0n vs. EL_f0a	56.5% 13/23	62.5% 15/24	29.2% 7/24	75.0% 18/24	55.8% 53/93

Table 2.2: Number and percentage of responses showing preference for the first token listed in each paired comparison (* $p < .01$).

A summary of the overall results obtained using the PC and VAS procedures across all four speaker-sentence conditions is shown in Table 2.3. Note that speech types judged to be closer to normal speech received higher PC scale values and lower VAS values. The rankings of the speech types by the two scaling procedures were identical. EL speech with amplitude-modulated F0 ("EL_f0a") was judged to sound better than EL speech with constant F0 ("EL_S"), but not quite as good as EL speech produced with the pre-laryngectomy F0 contour ("EL_f0n").

	PC		VAS			
Speech Type	Rank	Scale Value	Rank	Rating	Sm.	N
EL_f0n	1	1.63	1	6.5	0.17	117
EL_f0a	2	1.37	2	6.9	0.18	107
EL_S	3	0.0	3	7.3	0.09	13

Table 2.3: *Overall paired comparison and visual analog scale values*

2.3. Discussion

In this chapter, an approach for amplitude-based control of F0 in EL speech was developed and its impact on the quality of EL speech was examined. The approach utilized the positive linear relationship that was observed between F0 and amplitude in the pre-laryngectomy speech of EL users. The results of both PC and VAS experiments demonstrated that EL speech with amplitude-based F0-modulation was judged to sound more natural than EL speech with constant F0, thus lending preliminary support for using this simple linear relationship to compute an F0 contour for EL speech. Furthermore, analysis of the PC data using the binomial testing showed that there was no significant preference for the pre-laryngectomy F0 contour over amplitude-based F0-modulation implying that the listeners found these two types of stimuli relatively similar to each other. The scale values computed by analyzing the PC data also indicate that the perceptual distance between these two types of stimuli was relatively small. Compared to previously implemented F0 control methods using a finger-controlled button (Choi et al., 2001; Galyas, et al., 1982; Kikuchi & Kasuya, 2004; Takahashi et al, 2005; Tru-Tone) or stoma air pressure measurements (Sekey & Hanson, 1982; Uemi et al., 1994), the proposed F0 control scheme does not require access to alternative signals or control sources and may not require the extensive experience or training. Furthermore, this approach has the potential to be implemented with relative ease in close to real-time using a prototype (portable) DSP-based hardware platform. Possible configurations will be discussed in Chapter 6.

It must be noted, however, that this study was restricted to the improvement of

the naturalness of declarative sentences. As described in chapter 1, the F0 contour is important not only for the perceived naturalness of the EL speech but also for communicating linguistic contrasts such as intonation (e.g., declarative vs. interrogative) and contrastive stress. For example, interrogative sentences are associated with a maximal rise in F0 at the terminal portion of the utterance, while declarative versions are associated with a fall in the F0 during the terminal portion (Atkinson, 1973, 1976; Eady & Cooper, 1986). It has been also shown that stressed words have higher F0 values than when they are unstressed (O'Shaughnessy, 1979). Furthermore, in stress-accent languages, such as American English and Dutch, stress and accent were separate linguistic constructs and both have unique phonetic correlates (Okobi, 2006; Sluijter, 1995; Sluijter, Heuven, & Pacilly, 1997). More specifically, in these languages, F0 movement and overall intensity are acoustic correlates of pitch accents but not of stress, which is characterized by the longer duration and high-frequency emphasis. The next two chapters investigated the capabilities of amplitude-based control of F0 in different prosodic contexts by including sentences that were specially designed to vary intonation and stress patterns.

Chapter 3

Prosodic Control in F0 modified EL Speech: Intonation and Contrastive Stress

3.1. Introduction

3.1.1. General goals of the study

Based on the results of chapter 2, the study described in this chapter more fully evaluated the algorithms for F0 control. More specifically, formal perceptual evaluations of the F0-modified EL speech were conducted to investigate the prosodic control abilities of amplitude-based F0 control. Prosody refers to aspects of the speech signal that mark stress, rhythm, intonation, and pause structure (Lehiste, 1976). Acoustic parameters associated with prosody include F0, amplitude, duration, and segmental quality (Cooper, Eady, & Mueller, 1985; Eady & Cooper, 1986; Fry, 1958; Morton & Jassem, 1965; Shattuck-Hufnagel & Turk, 1996). Prosodic cues supplement the linguistic structure of the spoken message (Kent & Read, 2001), indicate the speaker's emotional state (Williams & Stevens, 1972), and are important for distinguishing between grammatical forms such as questions and statements (cf. Eady & Cooper, 1986).

3.1.2. Prosody in EL speech

As mentioned in Chapter one, a series of studies have investigated prosody in EL speakers. The prosodic functions that were investigated concerned perception and production of noun-verb contrasts ("OBject" versus "obJECT") (Gandour, Weinberg, & Garzione, 1983; Gandour, Weinberg, & Petty, 1986), minimally distinguished noun compounds and noun phrases ("BLACKboard" versus "black Board") (Gandour, Weinberg, & Kosowsky, 1982; Gandour & Weinberg, 1986), as well as contrasts on sentence level (question versus statements, and contrastive stress) (Gandour & Weinberg, 1982, 1983, 1984). The latter contrasts concerned sentences such as "Bev loves Bob", which could be produced as question or statement, and in which either name could be contrasted ("BEV" versus "Bev", and "BOB" versus "Bob").

Group	Intonation	Contrastive stress	Lexical Stress	Syntactic Stress
Normal	99.7	97.7	94.4	98.8
EL	54.0	79.9	82.8	81.8

Table 3.1: Mean percentage accuracy of identification of stress and intonation contrasts in normal and EL speech (from Gandour, Weinberg, & Kosowsky, 1982).

The summary of the results of these perceptual experiments is shown in Table 3.1., which revealed that listeners were able to identify the intended stress contrasts with relatively high accuracy in the EL speakers even without F0 variations. Acoustic analysis further showed that alaryngeal speakers using the EL marked contrastive stress effectively by increasing the duration of stressed syllables and by increasing the duration of pauses adjacent to stressed syllables (Gandour & Weinberg, 1984). On the other hand, Gandour and Weinberg (1983) showed that without F0 variation (F0 was constant throughout the utterance), it was impossible to communicate the difference between

question and statement, emphasizing the critical role of F0 as an acoustic cue for the perception of intonation.

3.1.3. Current study

The first goal of the current study was set to determine the impacts of amplitude-based control of F0 on the ability of EL users to produce linguistically meaningful contrasts on sentence level. More specifically, the ability of the F0 contour, as derived from the linear relationship of amplitude and F0, to convey the distinction between question and statement intonation was evaluated using a perceptual identification task. We also explored whether the amplitude control of F0 could further increase the ability of EL users to produce linguistically meaningful contrastive stress. F0 control scheme which adversely affects the communication of linguistic contrasts may not be desirable even though the same control scheme can improve the overall naturalness of the EL speech. Furthermore, determining the exact feasibility and limitation from different perspectives may suggest the way in which the algorithm can be improved and help develop an efficient training protocol in using the proposed control scheme.

Considering the different mechanics in speech production between EL speech and normal speech, however, several limitations are expected for the amplitude-based F0 control. In normal speech, voice production is dependent on the finely balanced relationship between the laryngeal configuration and respiration. The similarity of the F0 and amplitude patterns observed in normal speech in Chapter two can be explained by the fact that an increase in subglottal pressure increases the frequency of vocal cord vibration as well as the pressure of the sound wave (Fant, 1970). Intonation may be correlated with

particular patterns of change in subglottal pressure and this would give corresponding changes in both F0 and amplitude of the speech wave.

In EL speech, however, the amplitude of voicing source is essentially kept constant, so the amplitude fluctuation in EL speech output wave may be largely due to the supraglottal movement. As the quality of the sound changes from phoneme to phoneme, amplitude varies depending on the degree of mouth opening which is related to the segmental context. Producing linguistic contrasts using the amplitude-based F0 control is, therefore, expected to be more difficult compared to normal speech production, where the voicing source amplitude and the segmental aspects could be controlled relatively independently. For example, for the same segmental contexts, it was expected to be difficult to produce higher F0 values for stressed word compared to unstressed words in order to communicate the contrastive stress with amplitude-based F0 control. Communicating question intonation would be also limited, because the amplitude-based F0 contour is expected to fall at the end of utterances irrespective of the prosodic contexts, although F0 may need to be raised at the end in order to communicate question.

The second specific aim of this study was to evaluate the ability to convey linguistic contrasts with amplitude-based F0 control in comparison with the other F0 control scheme where the F0 is manually controlled (Tru-Tone, Griffin laboratories). In this type of device, a pressure sensor is built into a push button, and the F0 is controlled by the force (finger pressure) with which this button is pressed. In these products, there is a direct correspondence between operating amount and F0, so the F0 can be varied as desired with increasing practice. However, there have been no objective data examining the efficiency of manual F0 control approach in terms of communicating linguistic

contrasts and comparing different control schemes may further delineate the advantages and disadvantages of different control schemes.

In the manual control scheme, it was expected to see the differences in performance depending on the speaker's skill to control the finger pressure and achieve the desired F0 contour for a particular prosodic context. It was also expected to see the dependence of performance on particular types of sentences. For example, difficulties in providing intonational contrasts are expected, because vibration On/Off is determined by a threshold value which is set relative to the operating amount and speech cannot be started or ended with a desired F0 frequency.

In order to investigate the specific goals described in this section, we collected speech samples from normal four speakers, prepared and manipulated the F0 contours of utterances, and administered a perceptual listening task completed by normal hearing listeners using original and manipulated utterances as stimuli.

3.2. Methods

3.2.1. Sentence materials

The vocal tasks were similar to that used by Gandour and Weinberg (1983), and consisted of sentence quadruplets containing two declarative sentences and two interrogative sentences, with the location of contrastive stress differing within the statement/question pairs (Table 3.2). The four versions of each sentence were thus identical in their segmental composition and differed only with respect to where they

contained a focused word (on the initial or final word) and whether the intonation pattern was statement or question.

Version	Sentence type	Focus position	S1s1	s2
A	Statement	Initial	***BEV** loves Bob.*	***WE** were away.*
B	Statement	Final	*Bev loves **BOB.***	*We were **AWAY.***
C	Question	Initial	***BEV** loves Bob?*	***WE** were away?*
D	Question	Final	*Bev loves **BOB**?*	*We were **AWAY**?*

Table 3.2 Vocal tasks recorded that vary intonation and contrastive stress.

The sentences were short (three words) so they could be easily spoken in a single breath, to prevent the speakers from inserting pauses in the speech and to ensure they maintain the closed glottis condition. The first sentence ("Bev loves Bob") was chosen because it (a) is composed solely of monosyllabic words, a feature which eliminates word-level stress effects; (b) contains only voiced consonants; and (c) facilitates results comparisons with the data from previous studies (Atkinson, 1973, 1976; Gandour and Weinberg, 1982, 1983, 1984). The other sentence was chosen to have only vowels and semivowels, to make production of the EL speech as easy as possible.

3.2.2. Participants and recording procedures

The participants were four speech-language pathologists (2 female (f1, f2) and 2 male speakers (m1, m2)) who work with laryngectomy patients and are very experienced in training patients to optimize EL use. Vocal tasks were recorded digitally onto computer hard disk (Fs = 48 kHz) while the participants were seated in a sound-isolated booth. The recordings were calibrated for sound pressure level (SPL). The microphone was placed about one inch from the right corner of the speaker's mouth for EL speech

(EL device was placed against the left side of the neck) and about six inch from the center of the speaker's mouth for normal voice recording.

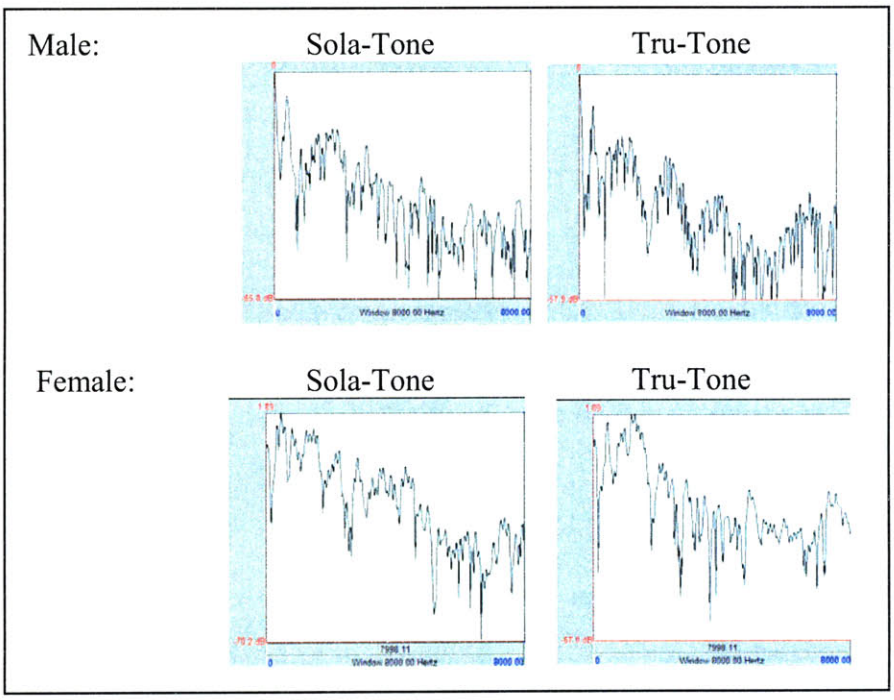


Figure 3.1: Spectra of voicing sources directly recorded from Sola-Tone EL and Tru-Tone EL.

Each participant recorded a standard set of speech tasks under three different conditions. In the first condition, subjects used their natural voices. The second condition involved producing the speech material using a Sola-Tone neck-placed EL (Griffin Laboratories) set at a constant F0. For the third condition, subjects produced the speech tasks using the manual control feature of the Tru-Tone EL (Griffin Laboratories) to vary F0. The acoustic and perceptual characteristics of voicing sources of Sola-Tone EL and Tru-Tone EL were similar to each other and the spectrums obtained directly from the two ELs (not filtered by the user's vocal tract) are shown in Figure 3.1. The difference between these two devices was thus confined to the presence of manual F0 control

feature for Tru-Tone EL. The recording order employed in this experiment allowed the speakers to become familiar with the vocal tasks before they were asked to manually control F0 with the EL. The speakers were instructed to hold their breath and maintain a closed glottis while talking with the EL, in order to approximate the modified anatomy of laryngectomy patients in which the lower airway was disconnected from the upper airway.

Dialogue 1:	Prompt:	Who loves Bob?
	Target:	BEV loves Bob. (SUE doesn't.)
Dialogue 2:	Prompt:	Sue loves Bob. Mary loves Bob. Bev loves Bob.
	Target:	BEV loves Bob? (Doesn't SUE?)
Dialogue 3:	Prompt:	Who does Bev love?
	Target:	Bev loves BOB. (Not DICK.)
Dialogue 4:	Prompt:	Bev loves John. Bev loves Charlie. Bev loves Bob.
	Target:	Bev loves BOB? (Not DICK?)
Dialogue 5:	Prompt:	Who was away?
	Target:	WE were away. (THEY weren't.)
Dialogue 6:	Prompt:	She was away. He was away. We were away.
	Target:	WE were away? (Weren't THEY?)
Dialogue 7:	Prompt:	Where were we?
	Target:	We were AWAY. (Not AT HOME.)
Dialogue 8:	Prompt:	We were away.
	Target:	We were AWAY? (Not AT HOME?)

Table 3.3: Sentence stimuli to investigate intonation and contrastive stress used in chapter three and chapter four. Target sentences were preceded by prompt sentences and contrastive-stressed words in target sentences were printed in large capital letters and underlined. In addition, a semantic context was enclosed within parentheses.

Each speaker was given a written set of instructions including a description of the semantic contexts in which the test sentences might occur (Table 3.3). Audio examples were also provided by a phonetically trained speaker with her normal voice, who is a senior research scientist in speech communication group at MIT. F0 contours provided by this speaker showed similar characteristics to those reported in the previous study by Gandour and Weinberg (1984). Examination of Figure 3.2 revealed that yes-no questions

were signaled consistently with a terminal rise in F0; statements were signaled with a terminal fall in F0. In the pair in which *Bev* was stressed, F0 fall slightly on *Bob* in statement tokens and rose sharply on *Bob* in question tokens. In the pair in which *Bob* was stressed, F0 rose and fall on *Bob* in statement tokens and rose sharply on *Bob* in question tokens.

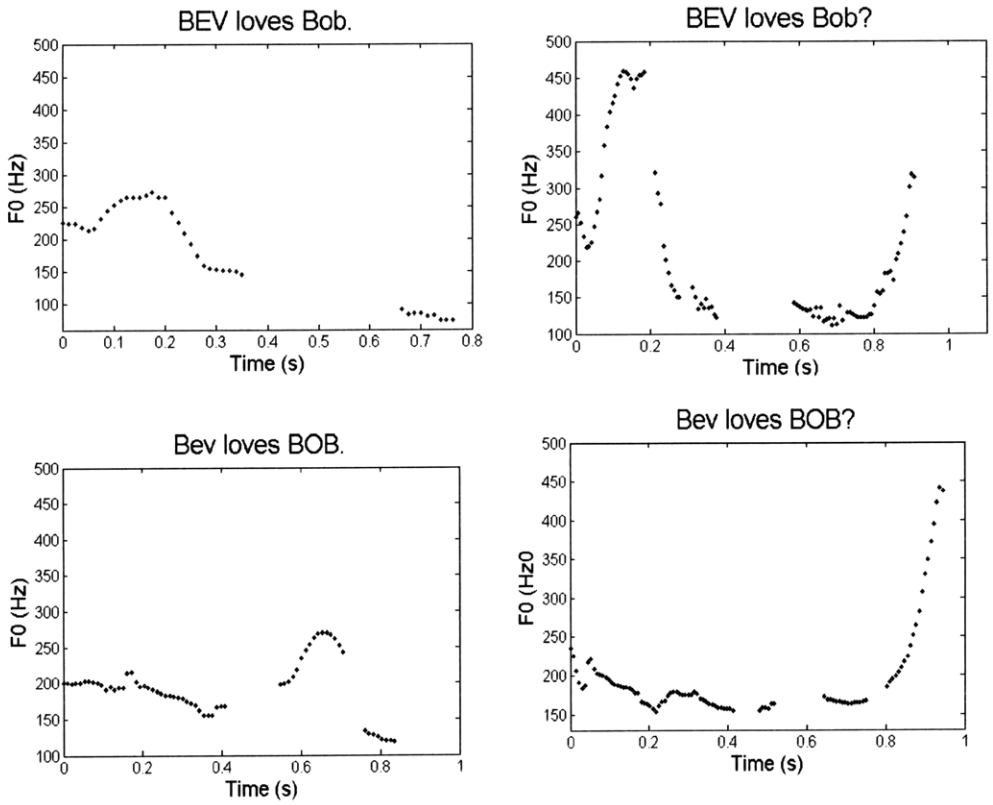


Figure 3.2: F0 contours of four sentences for sentence 1, each per sentence type by one reference normal female speaker. Words assigned contrastive stresses are indicated by capitalized letters.

The speaker was instructed to read the target sentences and place contrastive stress on the appropriate word via audio examples, and asked to read the sentence at his/her normal speaking rate (three times for normal voice and Sola-Tone EL speech with constant F0 condition and five to seven times for Tru-Tone manual F0 control condition).

3.2.3. F0 settings for EL devices

The F0 settings for the EL devices for this experiment are summarized in table 3.4. The baseline (lowest) F0 of the Tru-Tone EL could be adjusted from 50 Hz to 180 Hz. The device had an adjustable dynamic frequency range of 300 Hz from baseline that could be used to produce variable intonation. The baseline F0 frequency and dynamic F0 range for male speakers employed in this study were adjusted based on the previous study which showed improvement in intelligibility using manual F0 control compared to the constant F0 condition using the same types of EL devices (Tru-Tone, Griffin Laboratory) (Watson & Schlauch, 2009). The F0 settings for the female speakers were set to be approximately the double of those for the male speakers.

Condition	Male	Female
Sola-Tone EL : Constant F0	62.4	117.4
Tru-Tone EL: Manual F0 control	54.7 – 129.4	97.1 – 201.1

Table 3.4: F0 (Hz) settings of EL devices.

3.2.4. Generating the perceptual testing stimuli from the recorded material

For each speaker, three versions of each sentence were generated for use in the perceptual testing:

(a) EL speech with constant F0 (Sola-Tone, fixed F0)

(b) EL speech with F0 modulation based on the amplitude of the EL speech

(c) EL speech (Tru-Tone, varying F0) produced with manual F0 control.

3.2.4.1. Selecting the best tokens in EL speech with manual F0 control

The first step to generate stimuli for the perceptual experiment was to choose the best token for the manual F0 control condition. F0 contours produced with manual F0 control varied considerably across the repetitions for each sentence and we decided to select and include the best tokens in the perceptual experiment as opposed to selecting some random or average token in order to examine how prosodic contrasts could be conveyed by the manual F0 control as a preliminary analysis. The implication of this decision will be discussed later in the discussion section. Figure 3.3 shows the example of F0 contours produced with manual F0 control condition for sentence 1 (s1) in statement with stress on the initial word (version A) by one of the male speakers (m1). As we can see in this figure, the F0 contours showed great variability across the repetitions. The best token was selected by comparing the F0 contours produced by manual F0 control with those produced by the normal voice by the same speaker and listening to the corresponding audio files. Figure 3.3 on the right shows the F0 contour produced by the normal voice and Figure 3.4 shows the F0 contour of the selected token of EL speech with manual F0 control for the same sentence version (on the top-left panel) as well as other versions. The same procedure was followed for every sentence and every speaker. There were several sentences where none of the contours was close to the corresponding normal F0 contour and it was hard to select the best token. The examples of these cases will be described in chapter 4.

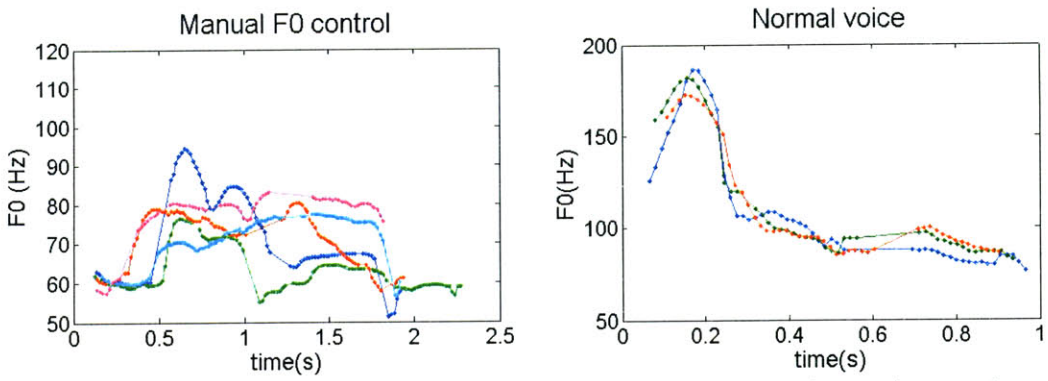


Figure 3.3: Left: F0 contours of five repetitions by one of the male speakers, m1, produced with Tru-Tone EL with manual F0 control. Right: F0 contours for three repetitions by one of the male speakers, m1, produced with his normal voice. The utterance was sentence 1 (s1) in statement with initial stress (*BEV loves bob.*)

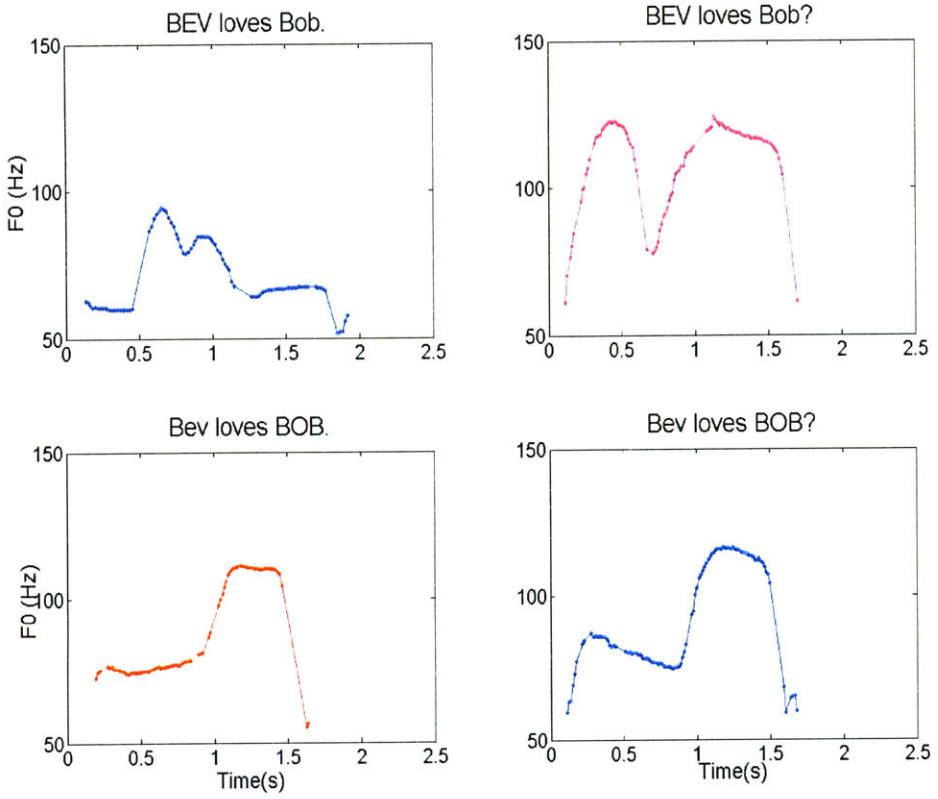


Figure 3.4: F0 contours of selected manual F0 tokens four sentences for sentence 1, each per sentence type by one of the male speakers, m1. Words assigned contrastive stresses are indicated by capitalized letters.

3.2.4.2. Generating amplitude-based F0 control tokens

The next step was to generate amplitude-based F0 control stimuli from the EL speech with constant F0 condition by linearly covarying the F0 based on the RMS amplitude variation as we developed in chapter 2.

$$F0 = k_1 + k_2 * RMS \quad (1)$$

This time, however, rather than using the regression coefficients between F0 and amplitude in the normal speech as we did in Chapter two, we determined the regression coefficients (k_1, k_2) for each speaker so that estimated F0 contours would be in the F0 range of selected manual F0 tokens. This is partly because the F0 and amplitude of the speech wave did not always vary in a similar way in normal speech wave (e.g., for question, the two curves were consistently different: F0 contour rises whereas amplitude contour falls at the end of the utterances). Another reason is to avoid any confounding factors related to the F0 range by having the same F0 range in amplitude-based and manual F0 control conditions. Table 3.5 shows the F0 range of the selected manual F0 tokens as well as the regression coefficients used to estimate F0 for each speaker.

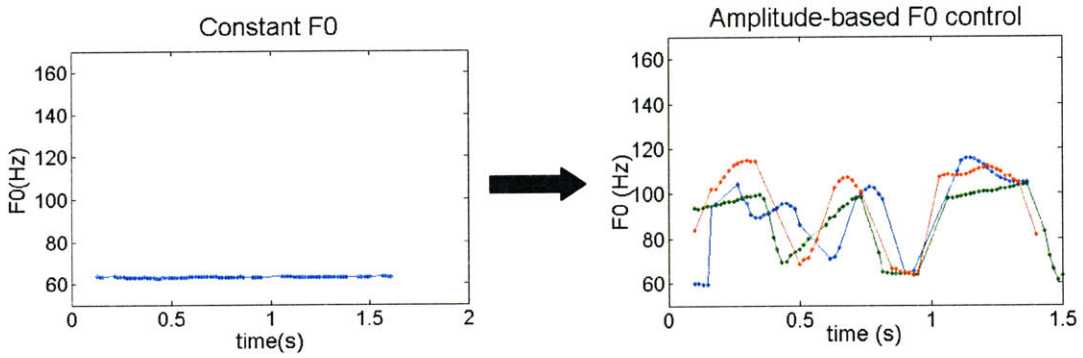
	m1	m2	f1	f2
F0 range (Hz)	503-135.3	47.9-108.9	99.5-186.3	97.9-177.8
k1	50.3	47.9	99.5	95.1
k2	578.9	429.8	309.6	915.8

Table 3.5: The F0 range (Hz) for selected manual F0 tokens and regression coefficients used to calculate the F0 contour based on the RMS amplitude contour for each speaker.

Examples of computed F0 contours based on the amplitude are shown in Figure 3.5. Compared to the F0 contours produced with manual F0 control, the F0 contours based on the amplitude were more consistent across repetitions. Therefore, from the three

realizations of each sentence, the second token was used for the experiment. If the quality of recording of the second token was not acceptable, either the first or the third tokens was used depending on the quality of recording.

BEV loves BOB.



Bev loves BOB.

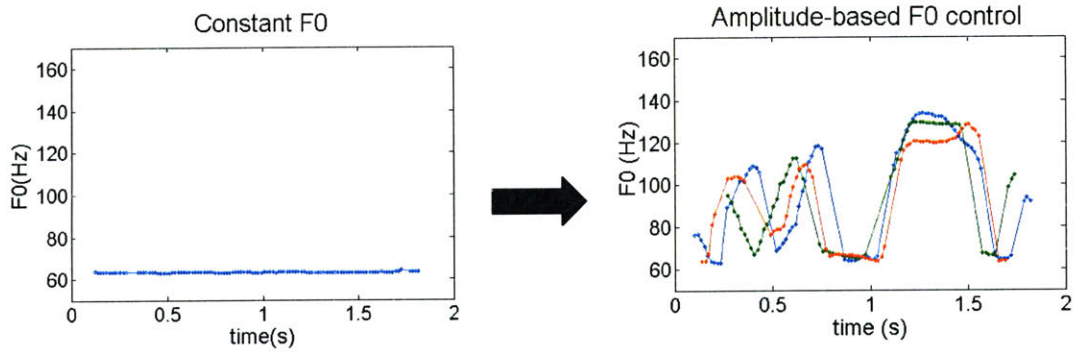


Figure 3.5: Examples of computed F0 contours based on the amplitude on the right panel and its original constant F0 on the left panel for sentence 1 in sentence with initial stress (BEV loves bob.) and final stress (Bev loves BOB) produced by speaker m1.

The data sets were manipulated using the Praat speech analysis software package (Boersma & Weenik, 2005) in order to modify F0 contours as computed and generate amplitude-based F0 tokens from the constant F0 tokens. The pitch synchronous overlap and add (PSOLA) technique was used to achieve F0 modulation without affecting the

tempo of the recording (Moulines & Charpentier, 1990). This process resulted in a set of 96 samples (2 sentences * 4 versions * 4 speakers * 3 conditions (manual F0 control, constant F0, and amplitude-based F0 control)). The intensity was equalized by setting the averaged RMS amplitude to be the same value for all sentences.

3.3. Assessment of Linguistic Contrast in F0 Modified EL Speech

3.3.1. Listeners

A total of 10 normal-hearing English speakers recruited at the MGH Institute of Health Professions served as listeners. Adequate hearing function was required of all listener participants. An audiometric screening evaluation was completed to ensure that average pure tone thresholds (at 500, 1000, 2000, and 4000 Hz) were at or below 25 dB HL in both ears.

3.3.2. Experimental procedures

Each sentence from the entire stimulus set (2 sentences * 4 versions * 3 conditions * 4 speakers = 96 sentences) was presented individually to a listener seated in a sound-isolated booth. A computer interface was used to present the stimuli and record listener responses. An intonation session was presented first, and then a contrastive stress session was presented in the next session. In the first session, for each vocalization, a selection button was displayed with a label of either "question" or "statement". Listeners were instructed to categorize each vocalization as either a question or a statement by

selecting the appropriate button. In the second session, the listener was asked if the current stimulus had sentence-initial or sentence-final stress, and s/he recorded that response onto a computer. When unsure, listeners were allowed to repeat the sentence once when they were not sure about the answer and asked to make their best possible judgment. There was no time limit for recording the response. The next sentence was presented one second after the listener pushed the button. To help the listener gauge his/her progress, a counter indicated the number of remaining trials in the experiment.

Before starting the actual experiments, a practice session was provided with normal voice recorded by the reference speaker. Feedback as to the correctness of their responses was given to the listeners during the practice sessions in order to familiarize them with the task and F0 contours in different prosodic contexts. After the practice sessions, the actual experiment was started. The sentence presentation order was random with respect to speaker, condition, location of stress, and sentence type. Each session (session for intonation and session for contrastive stress) was repeated to evaluate listener reliability. The listener was allowed to take five minutes of rest before going to start the second session. Feedback was not given to the subjects during the actual experiment.

3.3.3. Data analysis

The perceptual data were analyzed by means of a multifactor analysis of variance (ANOVA) procedure in order to assess the differences in listeners' perceptions of contrastive stress patterns and statement-question intonational patterns as a function of following main effects: (a) condition, (b) location of contrastive stress, (c) intonation, and (d) sentence for each speaker. The proportion of correct responses

calculated from individual listeners' responses to each speech token was transformed into angles, and the transformed data were then used in the analysis of variance. Arcsine transformations were used because homogeneity of variance could not be assumed when the observations in the analysis of variance were proportions (Winer, 1971). If there was any significant effect due to the condition, multiple comparison tests were further performed to determine which condition accounted for the main effect. The identification results were further divided into initial/final contrastive stress groups and statement/question groups to examine any possible correlation between accuracy in identifying intonation and contrastive stress location.

3.4. Results

3.4.1. Listener reliability

The intralistener agreement was evaluated using Pearson's *r* and was .79±.07 (the range was .65-.86) for contrastive stress and .76±.09 (.59-.90) for intonation. The results for the perception for contrastive stress will be described first followed by the results for intonation.

3.4.2. Perception of contrastive stress

The extent to which different F0 conditions in EL speech were able to distinguish contrastive stress was first assessed by determining the accuracy with which listeners identified the contrastive-stressed word. For each listener, classification accuracy score, a

proportion of correct responses out of 64 total utterances (32*2 sessions), was calculated for each of the three conditions. The average percentages of correct responses for each speaker as well as the overall correct percentages are shown in Table 3.6 and graphically represented in Figure 3.6 and Figure 3.7.

Contrastive Stress

Condition	m1	m2	f1	f2	Overall
Constant F0	63.7	68.1	71.2	62.5	66.4
Amplitude based F0	75.0	74.3	61.9	64.4	68.9
Manual F0 control	90.6	82.5	93.8	60.0	81.7

Table 3.6: Mean of percent correct depending on the condition for each speaker averaged over two sentences.

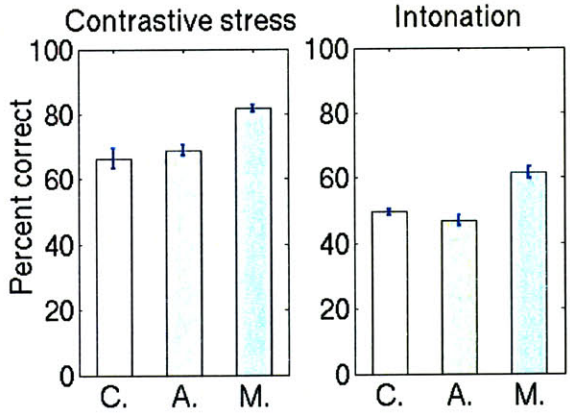


Figure 3.6: Percentage of correct responses averaged across two sentences and four speakers for each performance as a function of condition (C.: Constant F0, A.: Amplitude-based F0, and M.: Manual F0 control).

Overall, the manual F0 control achieved higher percentage of correct responses compared to the other two conditions. However, as Figure 3.6 illustrates, speakers performed the task with the varying degree of success for each condition. That is, some speakers were better than others at producing sentences with the required stress patterns for a particular F0 condition. For example, both of the two male speakers showed improvement in accuracy in the amplitude-based F0 condition compared to constant F0 condition, whereas the opposite result was obtained for one of the female speakers, f1.

We can also see that while the performance in the manual F0 control condition is higher than the other two conditions for three speakers (m1, m2, and f1), this was not true for speaker, f2.

The ranges of performance for the constant F0 and amplitude-based F0 conditions were much narrower across speakers than that in manual control F0 condition (Table 3.6). Specifically, the accuracy of identification for the constant F0 condition ranged from 62.5 – 71.2%. For the amplitude-based F0 condition, the accuracy of correct identification ranged from 61.9 – 75.0%; for the manual F0 condition, accuracy ranged from 60.0-93.8%.

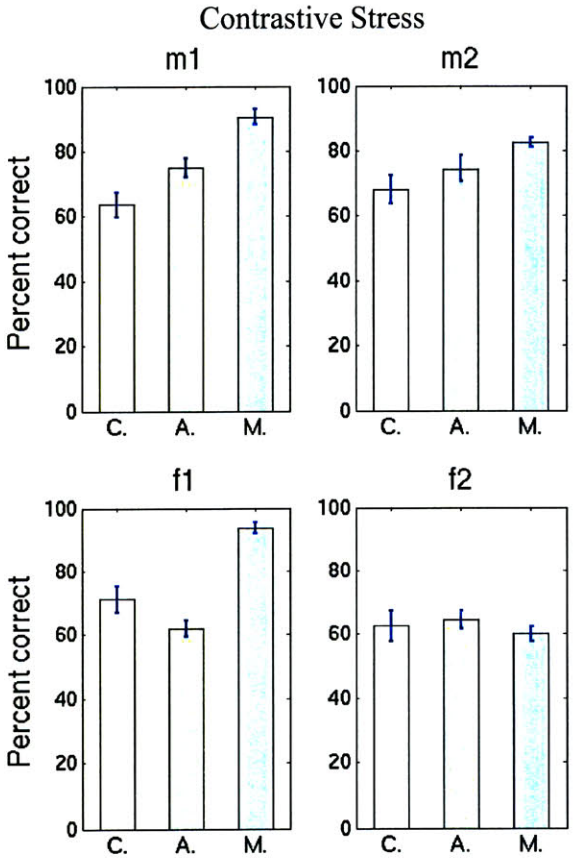


Figure 3.7: Percentage of correct responses of contrastive stress for each speaker averaged across two sentences as a function of condition.

3.4.3. Statistical analysis of perception of contrastive stress

Because there was a considerable variability across speakers, separate ANOVAs were performed on an individual speaker basis. The dependent variable was percentage correct answers (arcsine transformed) and independent variables were 1) three conditions (constant F0, amplitude-based F0, manual F0 control), 2) two sentences (Bev loves bob, We were away), 3) two stress locations (initial stress, final stress), and 4) two sentence types (statement, question).

The results of the statistical tests for all speakers are summarized in Table 3.7. The circle was entered for each factor with a significant effect. There was a significant difference among conditions for three speakers: m1 (F' (2, 216) = 16.46, MSE = 3.59, p < .001) , m2 ((2, 216) = 4.23, MSE = 1.03, p = .016, and f1 (F'(2 ,216) = 28.40 , MSE = 5.29 , p < .001)). Mean and standard error of transformed performance is shown in Figure 3.8. Post-hoc Pairwise comparison (Bonferroni corrected) further showed that for speaker m1, the percentage correct was significantly higher for manual F0 condition compared to constant F0 condition (p<.001) and amplitude-based F0 condition (p = 0.003) and amplitude-based F0 condition was significantly higher than constant F0 condition (p = 0.053). For speaker f1, the percentage correct was significantly higher for manual F0 condition compared to constant F0 condition (p<.001) and amplitude-based F0 condition (p<.001) but no significant difference between amplitude-based F0 and constant F0 conditions. For speaker m2, the percentage correct was significantly higher for manual F0 condition compared to constant F0 condition (p=.012).

Factors	m1	m2	f1	f2
Condition	O	O	O	
Sentence			O	
Stress location			O	O
Sentence type		O		
Condition * Stress location	O	O	O	O

Table 3.7: Results of ANOVAs of the stress performance for each speaker ($p< 0.001$).

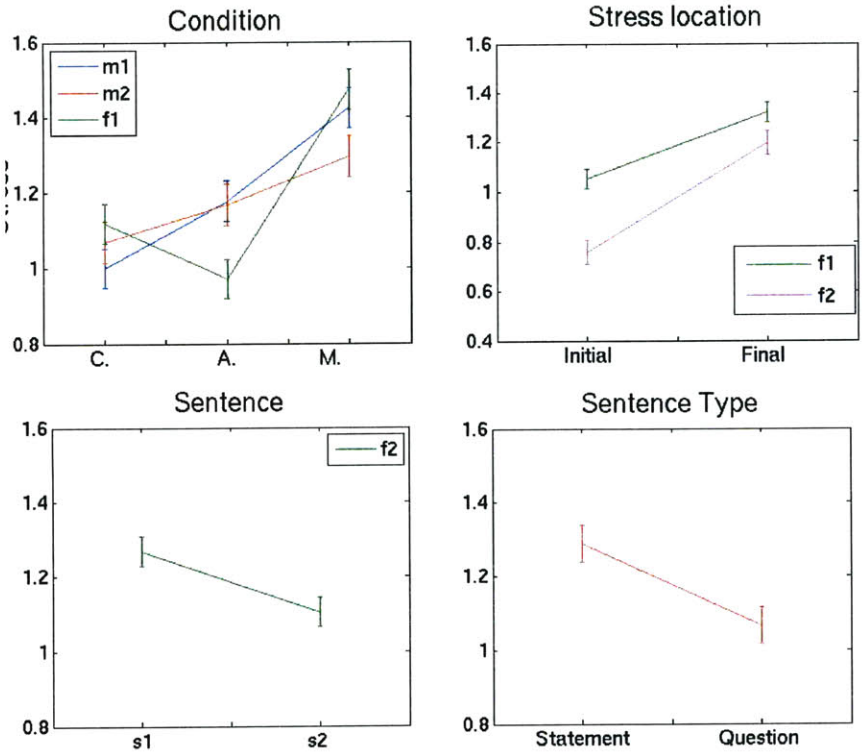


Figure 3.8: Mean and standard error of transformed percentage of correct responses of contrastive stress for each speaker.

The performance did not only depend on conditions, but also sentences, locations of the stress and sentence types (Figure, 3.8). Percentage correct was significantly higher for sentence 1 (80.8%) compared to sentence 2 (70.4%) for speaker f1 (F' (1, 216) =8.61, MSE =1.61, p = .004). The performance for final stress (84.2%) was significantly higher than the performance for initial stress (67.1%) for speaker f1 (F' (1, 216) =23.17, MSE =4.32, p <.001) and for speaker f2 (F' (1, 216) =43.39, MSE =9.56, p <.001) (initial stress=48.3%, final stress = 76.2%). The performance for statement (82.1%) was

significantly higher than the one for question (67.9%) for speaker m2 (F' (1, 216) =12.24, MSE =2.97, p =.001).

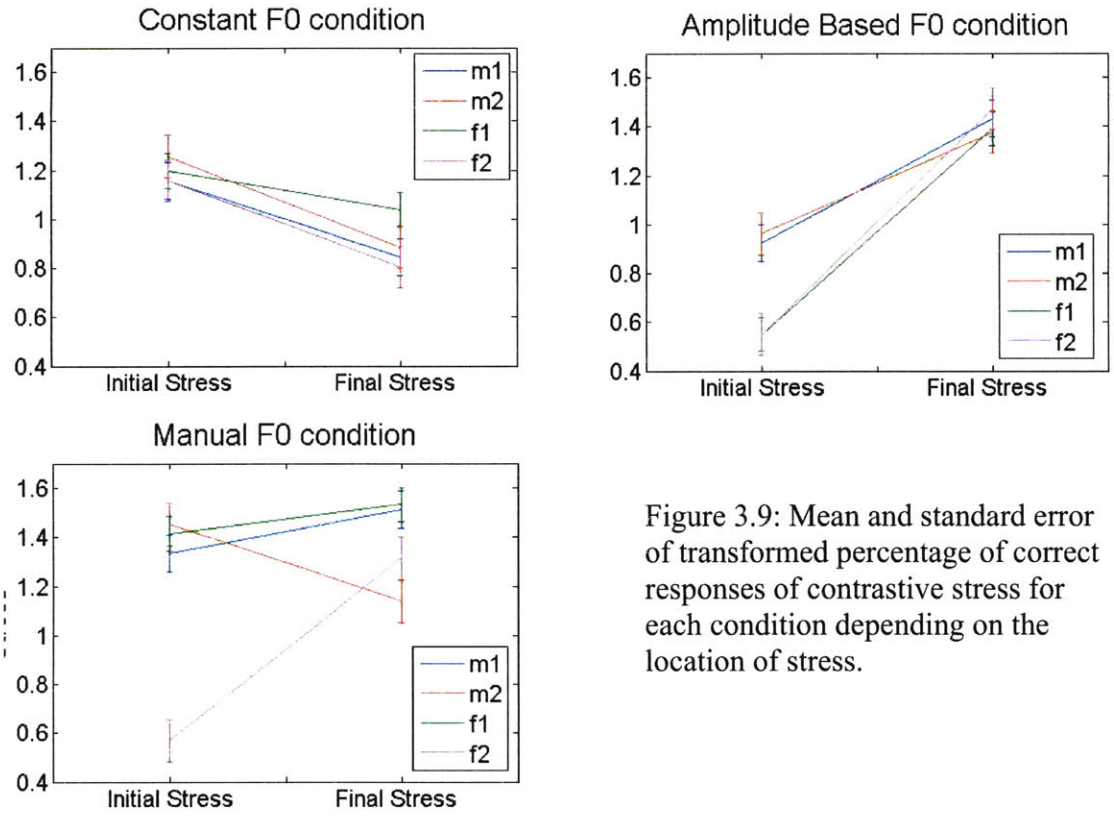


Figure 3.9: Mean and standard error of transformed percentage of correct responses of contrastive stress for each condition depending on the location of stress.

We also observed significant interactions between condition and stress location on an individual speaker level for all speakers: m1 (F' (2, 216) =15.75, MSE =3.44, p < .001), m2 (F' (2, 216) =9.53, MSE =2.31, p = .002, f1 (F' (2, 216) =28.71, MSE =5.35, p < .001, and f2 (F' (2, 216) =35.97, MSE =9.56, p < .001)). In other words, the performance of the stress perception depends on the condition and stress location. Figure 3.9 shows the performance depending on the stress location for each condition for each speaker. In constant F0 condition, listeners made more errors when listening to tokens with final stress rather than tokens with initial stress, although post-hoc tests did not show

statistically significant differences between initial and final stress. The percentage correct for initial stress was higher than that for final stress by 17.9 (±5.5) % (10~22.5%). The average percent correct was 74.9% for initial word and 57.0% for final word averaged across two sentences and four speakers. This characteristic did not depend on sentences, speakers, and sentence types.

In the amplitude-based F0 condition, the performance was generally higher for final stress than for initial stress by 42.7 ±15.9 % (initial, 47.5%; final, 90.3%) averaged across two sentences and all four speakers. The difference between initial and final stress was statistically significant for three out of four speakers (m1, f2, and f2) ($p < .001$). In the manual F0 control condition, the dependency of performance on stress location varies across speakers. The performance for final stress was significantly higher than initial stress condition for one female speaker (f2) ($p < .001$). There was no significant interaction for all the rest of the two-way and three-way interactions.

3.4.4. Perception of intonation

The extent to which different F0 conditions in EL speech were able to distinguish intonation was first assessed by determining the accuracy with which listeners identified the statement and question. The average percentages of correct responses for each speaker as well as the overall correct percentages are shown in Table 3.8 and graphically represented in Figure 3.6 and Figure 3.10.

Overall, the manual F0 control again achieved higher percentage of correct responses compared to the other two conditions. However, as Figure 3.10 illustrates,

there was some variability across speakers. The accuracy score was more or less similar across different F0 conditions for three out of four speakers (m1, m2, f1). For one of the female speakers (f2), there was a reduction in accuracy in amplitude-based condition compared to the constant F0 condition whereas manual F0 control achieved a relatively high level (81.9%) of performance.

Intonation

Condition	m1	m2	f1	f2	Overall
Constant F0	48.1 %	48.4 %	51.5 %	49.4 %	49.5 %
Amplitude based F0	53.1 %	51.9 %	45.6 %	36.9 %	46.9 %
Manual F0 control	56.9 %	45.6 %	61.3 %	81.9 %	61.4 %

Table 3.8: Mean of percent correct depending on the condition for each speaker averaged over two sentences.

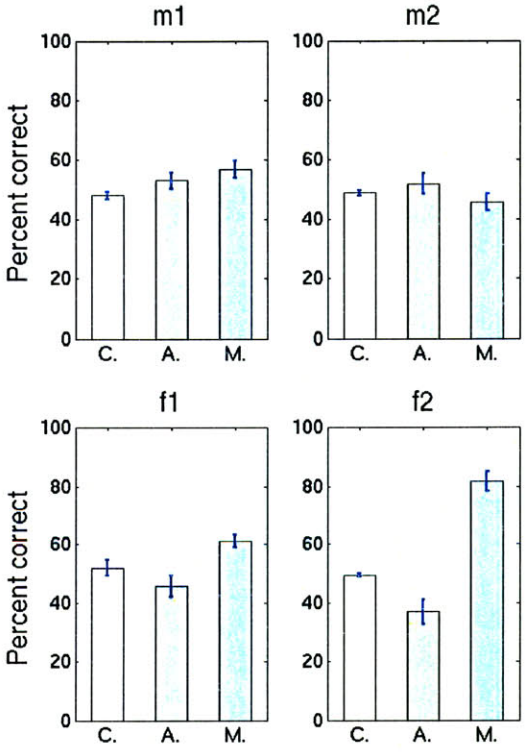


Figure 3.10: Percentage of correct responses of intonation for each speaker averaged across two sentences as a function of condition.

The ranges of performance for the constant F0 and amplitude-based F0 conditions

were again much narrower across speakers than those for manual control F0 condition (Table 3.8). The accuracy of identification for constant F0 condition ranged from 48.1 – 51.4%. For amplitude-based F0 condition, the accuracy of correct identification ranged from 36.9 – 53.1%; for manual F0 condition, accuracy ranged from 45.6- 81.9%.

3.4.5. Statistical analysis of perception of intonation

Separate ANOVAs were again performed on an individual speaker basis. The results are summarized in Table 3.9.a for main effects and Table 3.9.b for interactions. There was a significant difference among conditions for two female speakers: f1 (F' (2, 216) = 7.149, MSE = 1.221, p =.001) and f2 ((2, 216) = 57.645, MSE = 10.651, p < .001). Means and standard errors of transformed performance are shown in Figure 3.11. Post-hoc Pairwise comparison (Bonferroni corrected) further showed that for both speakers, the percentage correct of manual F0 condition was significantly higher than that of amplitude-based F0 condition (p<.001). For speaker f2, the listeners' performance for manual F0 control condition was significantly higher than the performance for constant F0 condition (p < .001).

Main factors	**m1**	**m2**	**f1**	**f2**
Conditions			O	O
Stress location				O
Sentence type	O	O	O	O

Table 3.9.a: Results of ANOVAs of the intonation performance for each speaker (main effects)(p<.001).

Interactions	**m1**	**m2**	**f1**	**f2**
Condition * Stress location		O		
Condition * Sentence type	O	O	O	O
Stress location * Sentence type	O	O	O	O
Condition * Stress location * Sentence type	O	O	O	
Condition * Sentence * Sentence type			O	

Table 3.9.b: Results of ANOVAs of the intonation performance for each speaker (interactions)(p< 0.001).

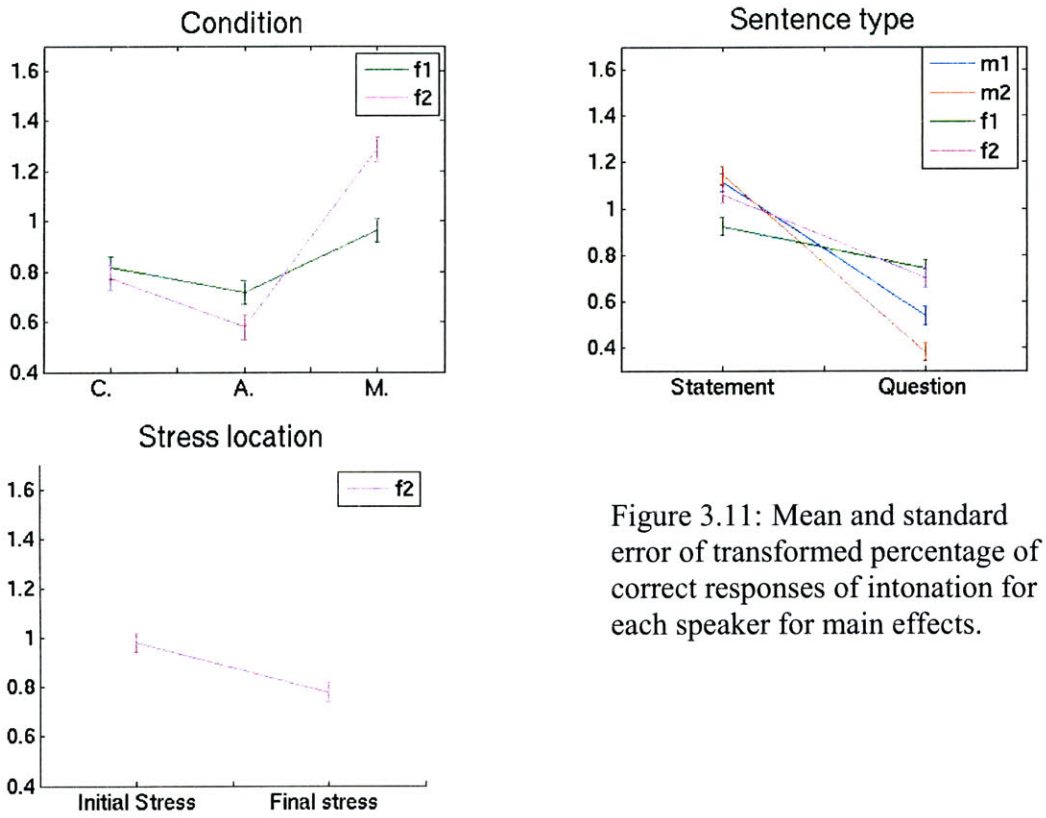


Figure 3.11: Mean and standard error of transformed percentage of correct responses of intonation for each speaker for main effects.

The performance depended not only on conditions, but also on location of stress and sentence types (Figure 3.11). The performance for initial stress (62.5%) was significantly higher than the performance for final stress (49.6%) for speaker f2 (F' (1, 216) =13.37, MSE =2.47, p <.001). The performance for statements was significantly higher than questions for all four speakers m1 (F' (1, 216) =106.93, MSE =19.883, p <.001), m2 (F' (1, 216) =201.64, MSE =35.18, p <.001), f1 (F' (1, 216) =11.79, MSE =2.02, p <.001), and f2 (F' (1, 216) =42.08, MSE =7.78, p <.001). The average percent correct across all speakers was 67.4% for statement and 34.9% for question.

As the Table 3.9.b illustrates, significant interactions between condition and sentence type on an individual speaker level were also found for all speakers: m1 (F' (1, 216) =61.08, MSE =11.36, p <.001), m2 (F' (1, 216) =54.27, MSE =9.47, p <.001), f1

(F' (1, 216) =90.41, MSE =15.35, p <.001), and f2 (F' (1, 216) =86.36, MSE =15.96, p <.001). In other words, the performance of the stress perception depends on the condition and sentence type. Figure 3.12 shows the performance of intonation depending on the sentence type for each condition and each speaker. In constant F0 condition, the error was mainly attributed to difficulty with classifying question tokens due to the fixed F0 contours across all speakers and sentences resulting in statistically poor performance in question (p<.001). In amplitude based F0 condition, the performance varied across speakers. One of the male speakers (m1) showed significantly higher performance for statement compared to question (p < 0.001). In manual F0 control conditions, the performance for question was higher than the performance for statement for female speakers (f1, f2), although the difference wasn't statistically significant.

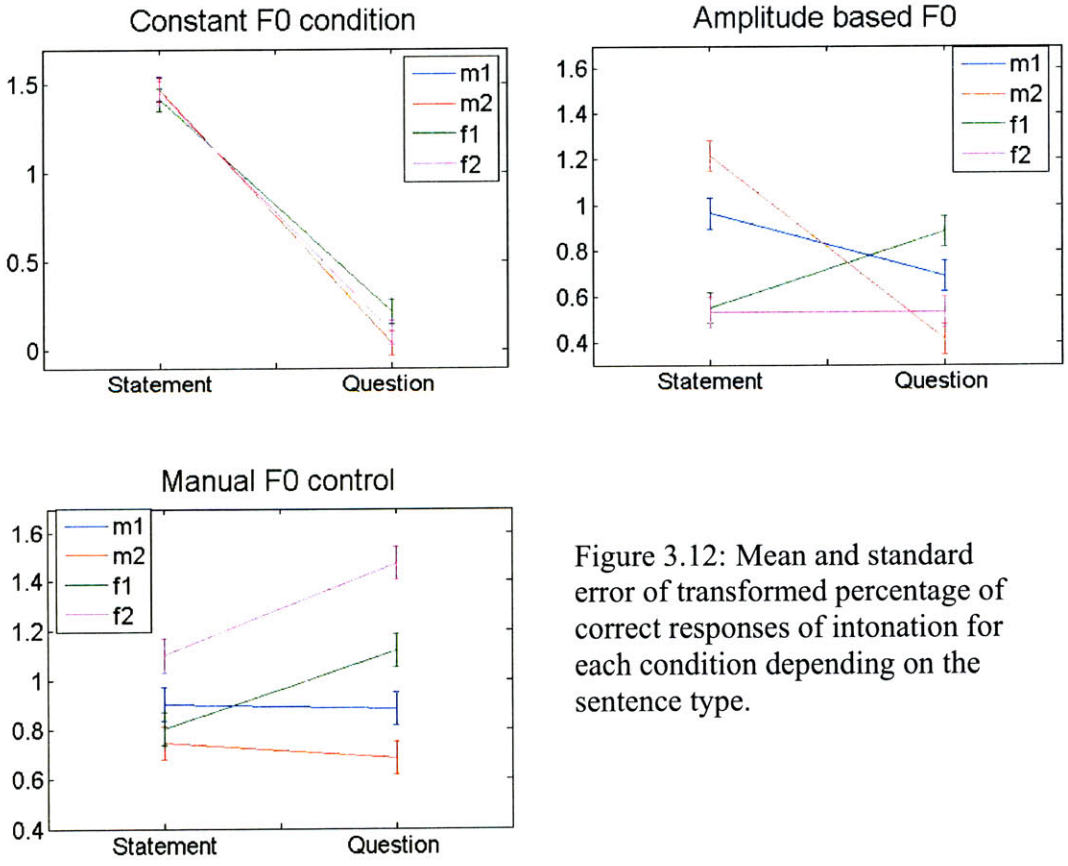


Figure 3.12: Mean and standard error of transformed percentage of correct responses of intonation for each condition depending on the sentence type.

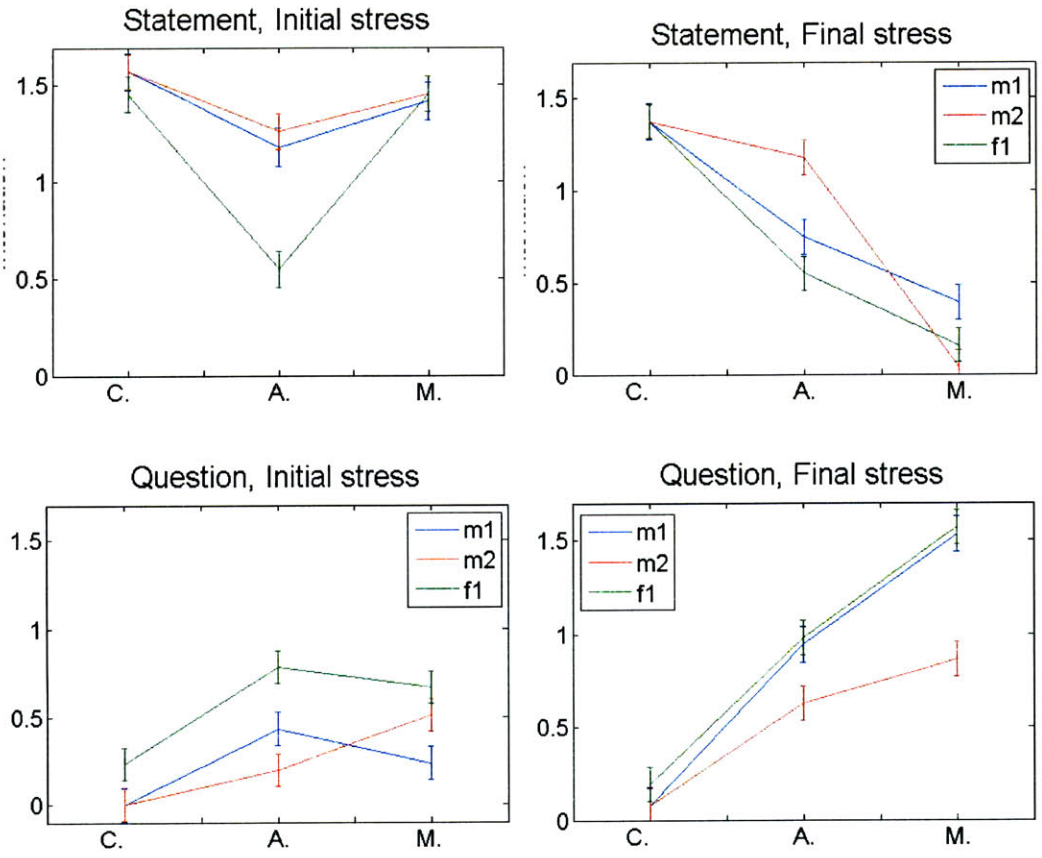


Figure 3.13: Mean and standard error of transformed percentage of correct responses of contrastive for each condition depending on the location of stress and sentence type.

Table 3.9.b also shows significant interactions between condition, stress location and sentence type for three out of four speakers: m1 (F' (1, 216) =29.14, MSE =5.42, p <.001), m2 (F' (1, 216) =18.44, MSE =3.22, p <.001), and f1 (F' (1, 216) =42.46, MSE =7.25, p <.001). Figure 3.13 shows the performance depending on the stress location and sentence type for each speaker. Post-hoc Pairwise comparisons showed that for statement with initial stress, the constant F0 and manual F0 condition were both significantly better than the amplitude-based F0 condition for one female speaker (f1) (p<0.001). For statement with final stress, manual F0 control condition generally showed significant

reduction in performance compared to constant F0 condition and amplitude-based F0 condition. Table 3.10 shows the summary of the Pairwise comparisons.

Pairwise comparison	**m1**	**m2**	**f1**
C. > A.	O		O
C. > M.	O	O	O
A. > M.		O	

Table 3.10: Results of Pairwise comparisons of the intonation performance for statement with stress on final word for each speaker ($p< 0.001$)

For question with initial stress, the performance was relatively low in general for all conditions for all speakers and the Pairwise analysis revealed that there was no statistically significant difference among different conditions. For question with final stress, the performance for manual F0 control and amplitude-based F0 conditions were significantly higher than constant F0 condition ($p<0.001$) for all three speakers. Although the performance for manual F0 condition was in general higher than amplitude-based F0 condition in general, there was no statistically significant difference between these two conditions.

3.5. Discussion

The goal of the current study was to determine the effects of controlling F0 based on the amplitude variation in EL speech on prosodic functions: signaling intonation and contrastive stress. First of all, the listener's responses for the constant F0 condition confirmed a set of findings reported by Gandour and Weinberg (1982,1983,1984) that it

was not possible to communicate the intonation distinction with constant F0 (averaged percent correct = 49.6%) illustrating the critical role of F0. The results for the contrastive stress perception also confirmed that the utterances could communicate the location of the stress to some extent (average percent correct = 66.4%) with no F0 cue available, implying that users manipulated other acoustic cues to communicate the contrastive stress.

Compared to the previous study by Gandour (1982), the percentage correct for contrastive stress in constant F0 condition is somewhat lower in this study (66.4% compared to 79.9 % in their study). This is presumably due to the differences in procedures employed in the perceptual experiments. In their study, a two-interval-forced-choice task was used: listeners had to indicate the order in which a spoken contrast was presented (e.g., first the initial stress and then the final stress, or the other way around). Thus listeners were presented with both contrasting utterances, so that they could make a comparison, which is relatively easier compared to normal speech communication and procedures used in the current study, where a listener had to identify the intention without directly comparing one intention with the other.

Compared to the result obtained in the constant F0 condition, overall results showed a slight improvement for perception of contrastive stress and slight degradation in perception of intonation in amplitude-based F0 control. The percentage correct for intonation and contrastive stress perception averaged across all sentences and speakers were 46.9% and 68.9%, respectively. As we expected, the performance of amplitude F0 control for contrastive stress seemed to be largely dependent on the segmental contexts. In this study, the amplitude of the vowels in the initial words (/e/ in "Bev" and /i/ in "We") are inherently smaller then that of vowels in the final words (/a:/ in "Bob" and /eI/

in "away") due to the difference in degree of mouth opening, which might have led to the higher correct percentage for sentences with final stress.

However, detailed analysis further revealed that the performance also depended on the speakers. For example, some improvement in stress perception in amplitude-based F0 control compared to the constant F0 condition was observed for two males. Although the main acoustic cues for contrastive stress and intonation is F0 contour, other acoustic variables such as duration, intensity, and spectral tilt have been documented as important acoustic variables for the perception of contrastive stress and intonation. Furthermore, some compensation has been observed in previous studies, in that a speaker with limited control over F0 might rely heavily on intensity or duration (Gandour & Weinberg, 1984; Slavin & Ferrand, 1995; McHenry, Reich, & Minifie, 1982). Thus it might be possible that the speakers who showed relatively higher performance in amplitude-based F0 condition tried to communicate the linguistic stress by either exaggerating acoustic cues other than F0 cues or systematically changing the degree of mouth opening to control the amplitude-based F0 depending on the presence of stress.

Another objective of this study was to compare the result with those obtained from EL with manual F0 control. Overall the listeners' accuracy in perceiving linguistic contrasts was higher for the manual F0 control condition compared to that of amplitude-based F0 control and constant F0 for both contrastive stress and intonation. Listeners were able to classify contrastive stress in EL speech with manual F0 control about 10 % higher than in the other two conditions.

However, it must be noted that, there was a great variability in F0 contours across repetitions in manual F0 control conditions, and only the best tokens were employed in

the perceptual experiments. On the other hand, the F0 contours estimated based on the amplitude were more consistent across the repetitions and the amplitude-based F0 control method has the potential for automatic control of F0. Furthermore, considering the typically advanced age of the EL users, it was expected to be relatively difficult to control F0 with manual pressure to convey the meaningful linguistic contrasts. Thus, the relative preference of patients for these two approaches can only be formally evaluated after the amplitude-based F0 control method has been implemented in real-time on the stand-alone, DSP-based hardware platform (detailed configuration will be described in chapter 5).

Detailed analysis of the performance in manual F0 control further showed that the percentage correct depended on the speaker, sentence, as well as on the sentence type and location of the stress. The variability of performance in manual F0 control condition would be partly due to the individual's skill in controlling finger pressure to achieve desired F0 contour. It was also shown that the performance for intonation was still very low, indicating some potential difficulties associated with this control scheme.

In order to investigate how the intonation and contrastive stress were communicated in different EL speech conditions and if there were problems, and what the possible acoustic basis was, EL speech used for the perceptual experiment were acoustically analyzed and relation between the performance and acoustic results were investigated in chapter 4.

Chapter 4

Acoustic Characteristics of Linguistic Contrasts in F0 Modified EL speech

4.1. Introduction

In this chapter, acoustical investigations of contrastive stress and intonation patterns produced in the different EL conditions were conducted in order to relate the acoustic data to the perceptual performance obtained in Chapter three. The two main questions of the present study included the following:

1. How did speakers use F0 and duration to signal the contrastive stress and question-statement contrast using EL in different F0 conditions?

2. How did these patterns differ from how normal speech signals the contrast?

Considering that the listeners perceived the location of stress to some extent in EL speech with constant F0, it was expected that the users controlled the durational cue to make the prosodic structure recognizable by the listener, as demonstrated in previous study (Gandour & Weinberg, 1984). Furthermore, speakers might compensate for the absence of F0 cues by exaggerating the temporal cues compared to normal speech condition (Gandour & Weinberg, 1984). The relatively higher performance for initial

stress compared to final stress was not, however, discussed in the previous literature and was examined in this study.

In amplitude-based F0 condition, the overall performance for contrastive stress location was more or less similar to what we observed in constant F0 condition. However, the distribution of the percentage correct as a function of the stress location was completely different. Listeners' accuracy in perception of stress was significantly higher for the final stress location than their accuracy for the initial stress location for three out of four speakers, possibly due to the segmental contents of sentences used in this study. In both sentences, the amplitude of speech wave was expected to be smaller in vowels in initial words compared to the amplitude in vowels in final words due to the differences in degree of mouth opening, leading to higher F0 values for final words and higher performance for final stress location irrespective of the prosodic context. We also tried to determine what the underlying acoustic basis was for the speaker variability observed in perception of contrastive stress.

Another important part of this study was to examine the underlying acoustic characteristics for the performance observed in manual F0 condition. In general, the tokens selected for manual F0 control conditions showed a high level of accuracy in perception of contrastive stress with mean of 81.7%. Thus, speakers might have manipulated acoustic cues important for contrastive stress perception relatively appropriately. It was also shown that the EL speech using the manual F0 control was able to signal the question-statement to some extent and human listeners could classify their productions with accuracy levels with a mean of 61.4%. It remained unclear, however, how speakers with manual F0 control conveyed the contrast when they succeeded. Did

they raise F0 throughout the utterance? Did they drop F0 instead of raising F0? Did they use the same F0 pattern as speakers with normal voice? Since the performance was still low, any underlying problems in communicating intonational contrast were examined in EL speech with manual F0 control.

In order to address these questions, the EL speech data set used in the perceptual experiments in Chapter three as well as a set of utterances produced with normal voices by the same speakers were examined. In this chapter, the method for acoustic analysis is first described in section two. Section three briefly summarizes the overall acoustic characteristics in normal speech. In section four, the results of acoustic analysis in EL speech are described in relation to the results of performance obtained in Chapter three.

4.2. Methods

4.2.1. Speakers and speech materials

The speakers were the same four speech-language pathologists (m1, m2, f1, f2) participated in the experiments described in Chapter three. In addition to the EL speech stimuli used in the perceptual experiments in Chapter three, a set of utterances produced with normal voices by the same speakers were used for the acoustic analysis and comparisons were made between different conditions in order to examine any differences in using the acoustic cues to communicate the contrastive stress and intonation depending on the conditions. The vocal task consists of sentence quadruplets containing two statements and two questions, with the location of contrastive stress differing within the statement/question pairs (Table 3.2). The speaker was instructed to place contrastive

stress on the appropriate word via audio examples and asked to read the sentence three times with normal voice. Of the three realizations of the utterances, the second token for each sentence was used for the analysis.

4.2.2. Acoustic analyses

In total, 128 utterances (96 EL utterances and 32 normal voice utterances) by the four speakers were acoustically analyzed. All utterances were sampled at 48 kHz and appropriately low pass filtered before resampled at 10 kHz. The acoustic cues (F0 and word duration) were investigated, using the Praat speech analysis program (Boersma & Weenink, 2005).

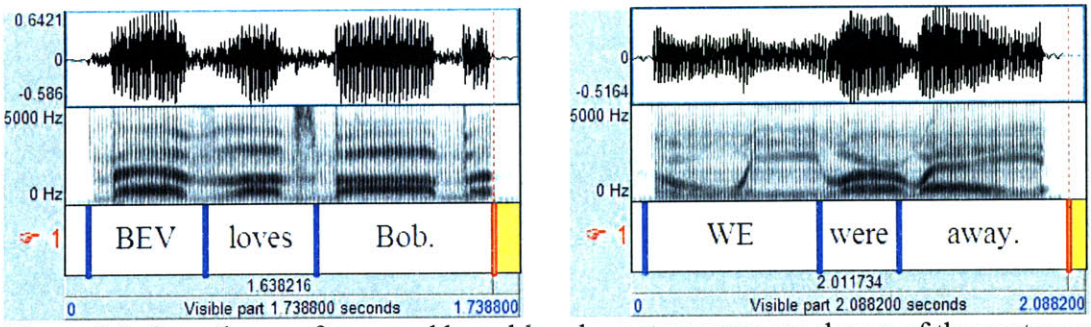


Figure 4.1: Speech waveforms and broad-band spectrograms are shown of the sentences produced by speaker m1 (EL speech with constant F0). Word boundaries are indicated by vertical lines drawn below the spectrograms.

4.2.2.1. Duration

The duration of each initial and final word was measured on the digitized waveform display using a computer-controlled cursor. Word boundaries were determined on the basis of combined audio-visual (spectrographic) information. Examples are shown in Figure 4.1. In the sentence, "Bev loves Bob", the boundaries between 'Bev' and 'loves' were characterized by a sudden change of spectrum which was due to an abrupt change in the articulation for /l/ in 'loves'. The boundaries between 'loves' and 'Bob'

were marked by the beginning of the closure for a labial stop consonant /b/ in 'Bob'. In the sentence 2, the boundary was defined to be the change in the second-formant frequency for /w/ and the change in the third-formant frequency for /r/ (Klatt, 1976; Perterson & Lehiste, 1960). More specifically, /w/ is characterized by very low first and second formant frequencies and /r/ is characterized by the very low frequency of the third formant.

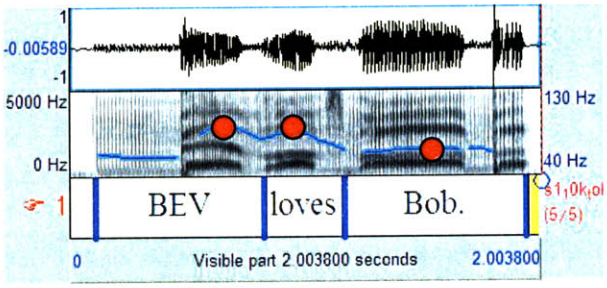


Figure 4.2: A speech waveform, spectrogram and F0 contour (in blue line) is shown of the sentence 1 produced by speaker m1 (EL speech with manual F0 control). F0 peaks for initial, second and final words are marked by red circles.

4.2.2.2. F0 peak

F0 values automatically generated by the Praat system often required manual correction because the pitch-tracking algorithm reported octave jumps that could not be verified auditorily. Manually adjusting the upper and lower F0 limits and frame duration parameters in Praat typically led to improved F0 tracking. These new F0 values were verified through visual and auditory inspection and confirmation using direct calculation of the F0 from the waveform. Praat-derived F0 values that continued to be judged to be errors were replaced by manually derived values obtained from the waveform.

The example of F0 contour generated by Praat system and measurements of F0 peaks are shown in Figure 4.2. The choice of the highest F0 value as the measurement point in each word was motivated by the fact that this point was very likely to show the effect of F0 heightening which was said to accompany increased focus on a word. It has been also

used as a measurement point in previous studies of F0 patterns (e.g., Cooper et al., 1985; Liberman & Pierrerhumbert, 1984).

4.2.3. Reliability of Acoustic Measures

Intrajudge reliability was assessed using a randomly selected sample of 12.5% of the EL speech with constant F0, EL speech with manual F0 control, and normal speech respectively. The total number of tokens examined was 12 (4 tokens x 3 conditions). Amplitude-based tokens were not sampled for the word duration measurement, since they had the same temporal patterns as the tokens in EL speech with constant F0 condition. Word-start and word-end points were manually relabeled 1 month after the original analysis. Intrajudge reliability of word duration measures across these two points in time was $r = .999$ ($M = 0.011$ s, $SD = 0.019$ s). Based on the new duration labels, all F0 peak values for this sample were recalculated. For the sample selected for constant F0, F0 peaks were calculated by looking at the corresponding samples of amplitude based F0 control. The correlation between the first and second measurements was 0.998. The mean difference between the first and second measurement was 2.1 Hz ($SD = 4.4$ Hz).

4.3. Acoustic Characteristics in Normal Speech

Before turning to the summary of the acoustic characteristics in normal speech, the speech rate (syllables/second) was first calculated in order to examine the differences in overall temporal characteristics between normal speech and EL speech.

4.3.1. Speech rate

Table 4.1 and Figure 4.3 show the speech rate for each speaker as a function of condition. The analysis of variance performed on this data yielded the following: condition, (F'(2,84) = 270.62, MSE = 14.68, (p<.001); speaker, (F'(3,84) = 11.82, MSE = 0.61, (p<0.001). The Condition x Speaker interaction was also significant (F'(6,84) = 6.31, MSE = 0.34, (p<.001). The Pairwise comparison (Bonferroni corrected) further showed that the speech rate was significantly slower in EL speech with constant F0 and manual F0 control conditions compared to normal voice condition for all speakers (p<.001). This observation was consistent with the previous study where the EL speech has been shown to be characterized by a very slow speaking rate (Bennet & Weinberg, 1973). Furthermore, we found significantly slower speech rate in EL speech with manual F0 control than in EL speech with constant F0 condition for speaker m1 (p<.001).

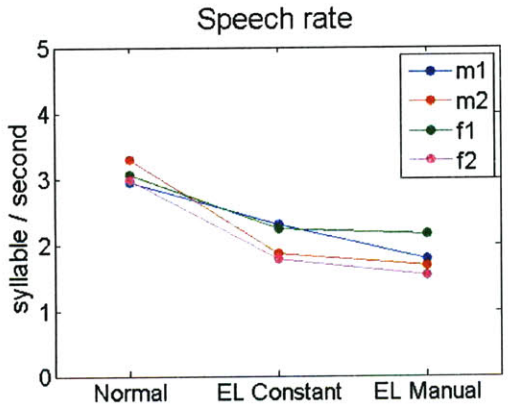


Figure 4.3: Mean speech rate (syllable/second) averaged over two sentences for each speaker as a function of the condition.

Condition	m1	m2	f1	f2	Overall
Normal speech	2.95 (0.28)	3.29 (0.40)	3.07 (0.24)	2.99 (0.26)	3.08 (0.32)
EL speech with constant F0	2.31 (0.23)	1.87 (0.09)	2.25 (0.19)	1.79 (0.09)	2.05 (0.28)
EL speech with manual F0	1.79 (0.11)	1.69 (0.15)	2.16 (0.27)	1.54 (0.27)	1.80 (0.31)

Table 4.1: Speech rate (syllable / second) averaged across two sentences for condition (values in parentheses are standard deviations).

4.3.2. Duration in normal speech

Results of the duration analysis for the sentences are presented in Table 4.2, where the means are arranged according to sentence versions and word positions. Each value in the table represents the mean for two sentences averaged across all four speakers. These results are presented graphically in Figure 4.4. It is clear from these results that, as expected, the mean duration of a word was greater when it was stressed (88.8 ms on average). The result of the ANOVA reveals that the difference in duration among the four sentence versions was significant for both word positions [for sentence-initial words, $F'(3, 28) = 15.71$, MSE = 21643.6, $p<.001$; for sentence-final words, $F'(3, 28) = 6.00$, MSE = 37144.1, $p=.003$].

Version	Sentence type	Focus position	Initial	Final
A	Statement	Initial	247.9 (31.0)	401.8 (106.1)
B	Statement	Final	190.6 (31.1)	526.8 (57.1)
C	Question	Initial	294.6 (55.9)	498.2 (88.9)
D	Question	Final	184.5 (21.3)	560.1 (48.4)

Table 4.2: Mean duration (ms) of words in normal speech averaged across two sentences produced by four speakers (values in parentheses are standard deviations).

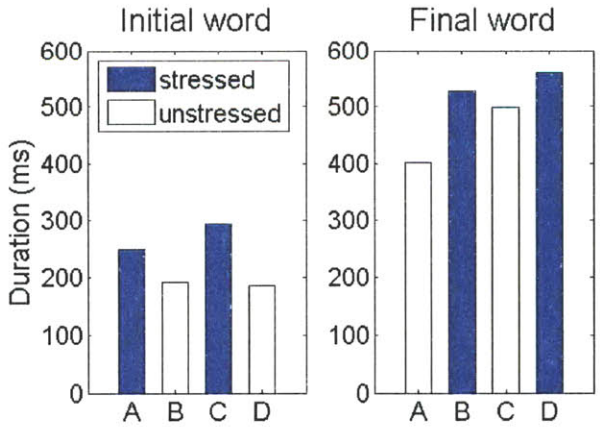


Figure 4.4: Mean duration in normal speech for the initial and final words, averaged across two sentences and four speakers. The sentence versions are shown in Table 4.2.

For words in sentence-initial positions, Pairwise comparisons (Bonferroni corrected) reveals that the two versions containing sentence-initial focus had significantly

greater values than those without focus in this position (A>B, p =.027; A>D, p =.012; C>B, D, p<.001). Further analysis of the data reveals that this pattern holds for two sentences and all four speakers. The present result also shows that the question/statement variable seems to exert no demonstrable effect on the duration of stressed and unstressed words in sentence-initial position.

For words in sentence-final position, the durations of the two stressed versions were significantly greater than those of the unstressed versions (i.e., B > A, p=.022; D >A, p =.002). However, there was no significant difference between the two stressed versions and version C (question with initial stress). For sentences with initial stress, although the difference was not significant, the duration of final word was somewhat lengthened in question (version C) than in statements (version A). The average difference was 96.4 ms.

4.3.3. F0 peak in normal speech

The results of the fundamental frequency analysis for the sentences in this study are presented in Table 4.3. Each value in the table represents the mean peak F0 for each word position averaged across two sentences for two male speakers and two female speakers. These results are graphically presented in Fig. 4.5.

	Male			Female		
Version	First	Second	Third	First	Second	Third
A	200.7 (43.6)	151.5 (44.9)	91.8 (6.9)	352.1 (53.4)	251.6 (49.2)	177.9 (14.8)
B	108.4 (6.6)	101.3 (11.9)	172.4 (34.5)	212.4 (52.1)	193.8 (15.7)	291.7 (33.0)
C	253.4 (36.1)	116.0 (48.8)	170.6 (33.6)	458.8 (148.3)	218.4 (22.9)	236.7 (38.6)
D	108.7 (3.2)	150.5 (1.4)	263.9 (7.2)	203.3 (11.1)	269.3 (14.2)	478.4 (70.4)

Table 4.3: Mean fundamental frequency (Hz) of each word in normal speech averaged across two sentences for two male and two female speakers (values in parentheses are standard deviations).

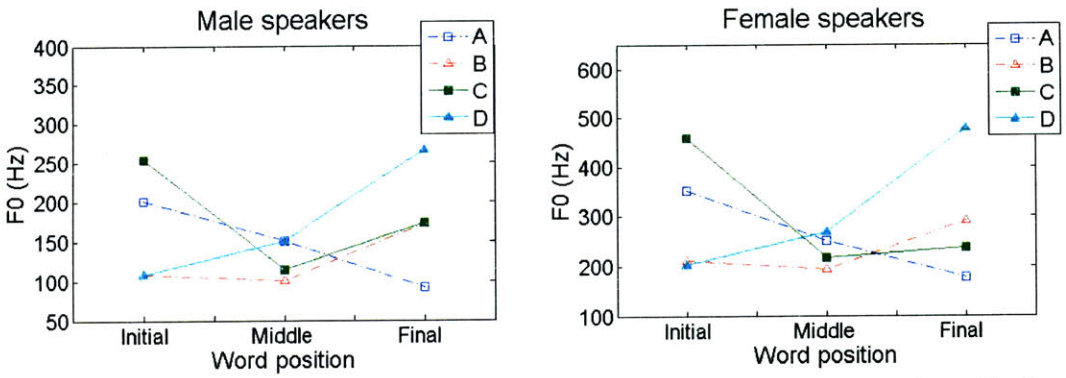


Figure 4.5: Average F0 peaks in normal speech for the four sentence versions. Each peak represents the average for two sentences spoken by two male and female speakers. The sentence versions are described in Table 4.2.

ANOVAs were calculated at each of the three word positions to determine whether there were significant differences in the F0 patterns for the four sentence versions. The results of these ANOVAs indicate significant differences in F0 at the initial and final word position for both male and female speakers {male initial word, $F'(3, 12) = 25.26$, MSE= 20560.96, $p < .001$, male final word, $F'(3, 12) = 32.80$, MSE = 19812.19, $p < .001$, female initial word, $F'(3, 12) = 45.35$, MSE= 59692.75, $p < .001$, female final word, $F'(3, 12) = 34.89$, MSE= 67665.67, $p < .001$}. There was no significant difference, however, among the F0 values of the four sentence versions at the second word position for both male and female speakers.

4.3.3.1. Statements: initial stress (version A) vs. final stress (version B)

Version A, with sentence-initial focus, had a significantly higher F0 value on the initial word than did the version B for both male speakers ($p = .004$) and female speakers ($p=.001$). At the initial word position, version A (with sentence-initial focus) increased to a mean value that is 92.3 Hz and 139.7 Hz higher than version B for male and female speakers respectively. This pattern holds for two sentences and all four speakers.

On the other hand, the final, stressed word of version B had a significantly higher F0 value than version A for both male (p =.003) and female speakers (p =.02). Version B had an average rise of 61.3 Hz and 96.4 Hz for male and female speakers between the second and third words, presumably due to the presence of stress on the last word of the sentences and increased to a mean value that was 80.6 Hz and 113.8 Hz higher than version A for male and female speakers respectively. This pattern holds for two sentences and all four speakers.

4.3.3.2. Question: initial stress (version C) vs. final stress (version D)

We also observed a difference in the F0 patterns of the two question versions at the initial and final word positions. For the initial word, Pairwise comparisons among the mean F0 values reveal that the frequency of version C (initial stress) was significantly higher than version D (final stress) for both male and female speakers (p<.001). Again this pattern of results was consistent for all two sentences and all four speakers. At the final word location, version D was significantly higher than version C for both male speakers (p=.001) and female speakers (p<.001).

4.4.3.3. Question vs. statement

The F0 patterns presented in Fig 4.5 show the question version in this study (the contours connected with solid lines) to have generally higher F0 values than the corresponding statement versions (the contours connected with dotted lines), with a particularly high value at the end of the sentence. For male speakers, Pairwise comparisons among the means at the final word position reveal that the questions had

significantly higher F0 values than the statements (C vs. A (p=.004), D vs. B (p=.001)). There was also a significant difference among the two question versions at this sentence position (p=.001) (D vs. C). This pattern of F0 results holds for two sentences and two male speakers in this study. For female speakers, again the Pairwise comparisons among the means at the final word position show that the question had significantly higher F0 values than the statements (D vs. A, B) (p<.001). However, there was no significant difference between C vs. A, B. Again the F0 peaks was significantly higher in version D compared to versions C (p<.001)).

4.3.4 Summary of acoustic analysis in normal speech

As a summary, the results obtained in the analysis of normal speech are in agreement with the earlier studies (Cooper et al., 1985; Eady & Cooper, 1986) and confirm the previous finding that the placement of contrastive stress causes a significant durational increase and higher F0 peak on stressed word. The question intonation further introduces a significantly higher F0 peak on the final word.

4.4. Relationship between Listener Perception and Acoustic Characteristics in EL speech

Analyses were conducted to understand the relationship between listener perception and acoustic characteristics of the EL speech in order to better understand how speakers with different types of EL devices signal the prosodic contrast. Comparisons of acoustic characteristics between EL speech and normal speech were also made where necessary. In this section, the analysis for the contrastive stress will be described first, followed by the results of the analysis for intonation.

4.4.1. Contrastive stress

4.4.1.1 Constant F0 condition

In EL speech with constant F0 condition, we examined if there were durational cues to signal the contrastive stress. The results of the duration analysis are presented in Table 4.4. Figure 4.6 shows the average word duration for each version for initial and final words. As in the normal voice condition, the stressed word was generally longer than unstressed word. The average increase was 104.3ms across two sentences and four speakers. The result of analyses of variance revealed that the difference in duration among the four sentence versions was significant for sentence-initial words, F' (3, 28) = 4.85, MSE = 24404.56, p=.008, but not for sentence-final words. Pairwise comparisons further showed that, for initial words, stressed words in version A were significantly longer than unstressed words in version B and D (A>B, p=0.095; A>D, p =.043).

Version	Sentence type	Focus position	Initial	Final
A	Statement	Initial	439.2(93.0)	712.1 (128.7)
B	Statement	Final	348.5 (44.1)	818.7 (110.3)
C	Question	Initial	435.6 (49.7)	730.7 (94.8)
D	Question	Final	335.9 (84.0)	851.0 (114.2)

Table 4.4: Mean duration (ms) of words in EL speech constant F0 averaged across two sentences produced by four speakers (values in parentheses are standard deviations).

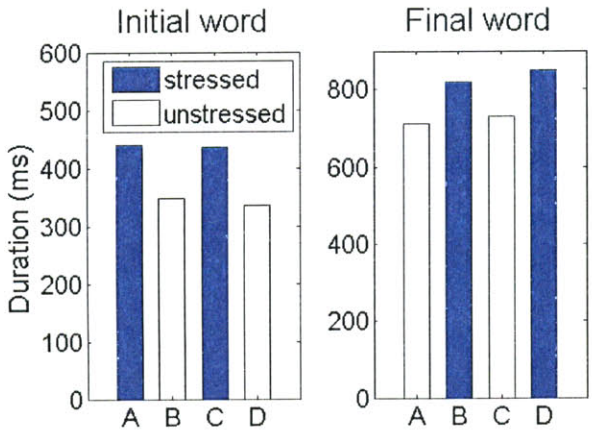


Figure 4.6: Mean duration in EL speech with constant F0 for the initial and final words, averaged across two sentences and four speakers. The sentence versions are shown in Table 4.2.

Condition	Initial word	Final word
Normal Voice	47.2 %	26.3 %
EL with Constant F0	31.7 %	16.6 %
EL with Manual F0	35.1 %	15.6 %

Table 4.5: Mean percentage increase in duration due to the contrastive stress averaged over two sentences and four speakers.

We also examined if there was any difference in durational patterns between normal speech and EL speech. One striking difference was the greater averaged durations found in EL speech. Word duration in constant F0 condition was longer than the corresponding word produced by the normal voice by 219.2 ms on average (the average difference was 158.6 ms for the initial word and 60.3ms for the final word). The results of analysis of variance revealed that the difference in word duration was significant for condition (normal speech, EL speech with constant F0, and EL speech with manual F0 control) ($F'(2,186) = 120.25$, MSE= 1754293.59, $p<.001$) and for stress location (initial stress and final stress) ($F'(1,186) = 333.29$, MSE = 4862096.214, $p < .001$). The

interaction between condition and stress location was also significant ($F'(2,186) = 4.320$, MSE= 63024.54, p =.015). Pairwise comparison revealed that the word duration in constant F0 condition was significantly longer than normal voice condition for both initial and final words (p<.001).

Lengthening in EL speech was not, however, confined to those that contained a contrastive stress. Table 4.5 shows the average percentage increase in duration due to the contrastive stress for the initial and final word for each condition. These results show that while EL speech was in general of longer duration than normal speech, the percentage increase in duration in EL speech was smaller than in normal speech and there was no temporal exaggeration present in the words containing stress that could be interpreted as a compensation for the missing F0.

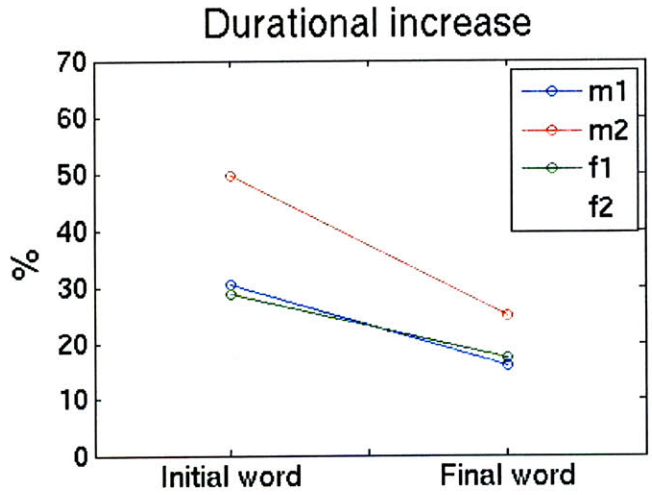


Figure 4.7: Mean percentage increase in word duration due to contrastive stress for each speaker as a function of word position. Each point represents the average percentage across two sentences for each speaker.

Table 4.5 also indicates that the words occurring at the end of a sentence were lengthened to a lesser degree than words at the sentence-initial positions. The percentage increase in duration for each word position for each speaker is shown in Figure 4.7. The

generally higher percentage correct in contrastive stress perception for initial stress compared to that for final stress (the average percent correct was 75.9% for initial word and 57.0% for final word) (Figure 3.9) would seem to be explicable on the basis of the difference in the amount of durational increase as a function of sentence position. In order to test this hypothesis, we calculated Spearman rank-correlation coefficients to determine whether a liner relationship exists between the mean percentage correct and the percent increase in duration due to the contrastive stress for each speaker and for each word position. The results indicate that there is a marginal positive correlation between the percentage increase in duration and the percentage correct ($r = 0.74$, $p = 0.036$) (Figure 4.8).

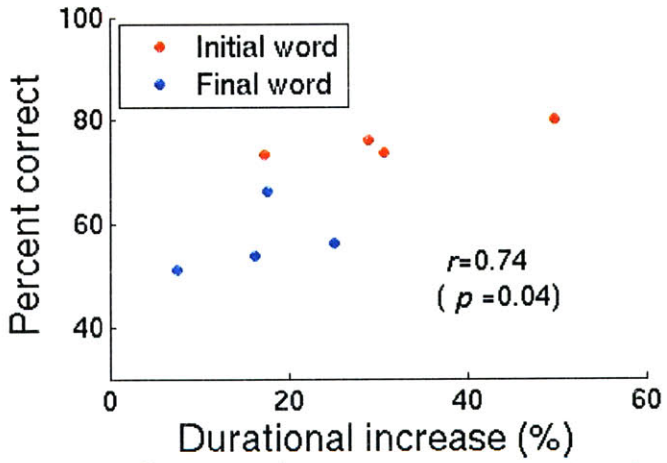


Figure 4.8: Percent correct for contrastive stress versus percentage increase in word duration. Each point represents the average percentage across two sentences for each speaker.

4.4.1.2 Manual F0 control condition

Results of the duration analysis are presented in Table 4.6 and graphically represented in Figure 4.9. In manual F0 control condition, the stressed word is longer than the unstressed word by 122.0 ms on average. However, there was no significant

difference in duration among the four sentence versions for both initial and final word in both male and female speakers.

Version	Sentence type	Focus position	Initial	Final
A	Statement	Initial	607.1 (190.4)	767.6 (154.5)
B	Statement	Final	418.1 (102.8)	920.1 (94.5)
C	Question	Initial	616.8 (178.6)	810.7 (133.9)
D	Question	Final	508.3 (195.9)	848.5 (152.2)

Table 4.6: Mean duration (ms) of words in EL speech with manual F0 control averaged across two sentences produced by four speakers (values in parentheses are standard deviations).

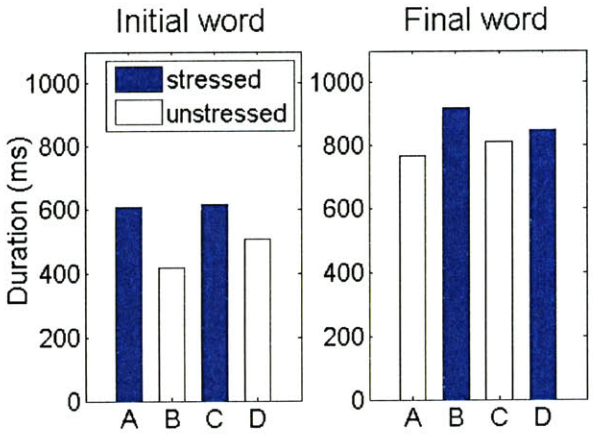


Figure 4.9: Mean duration in EL speech with manual F0 control for the initial and final words, averaged across two sentences and four speakers. The sentence versions are shown in Table 4.2.

Compared to the normal speech, the average duration in EL speech with manual F0 control was significantly longer for both initial words and final words ($p<.001$) (Pairwise comparisons of the ANOVA described in section 4.4.4.1). The word duration is longer than the corresponding words in normal speech by 324.1 ms on average. Furthermore, the word length in manual F0 control condition was significantly longer than the corresponding words in constant F0 condition for initial words ($p<.001$).

The results of the F0 peak analysis are presented in Table 4.7 and graphically represented in Figure 4.10. As in normal speech, the results of the ANOVAs indicate significant differences in F0 at the final word position for both male and female speakers

{male, $F'(3, 12) = 6.94$, MSE= 911.21, $p = 0.006$, female, $F'(3, 12) = 6.55$, MSE = 1225.69, $p = 0.007$}. In particular, pairwise comparisons showed that version A, with sentence-initial focus in statement had a significantly lower F0 value on the final word than did the other three versions (B, C, and D) (B>A, $p = 0.022$; C>A, $p = 0.023$; D>A, $p=0.011$) for male speakers. For female speakers, version A, with sentence-initial focus in statement had a significantly lower F0 value on the final word than did version D (question with final stress) ($p=0.005$).

	Male			Female		
	First	Second	Third	First	Second	Third
A	105.5 (9.5)	101.1 (5.4)	76.5 (12.6)	152.3 (22.5)	133.2 (15.6)	119.0 (10.9)
B	77.7 (4.3)	80.0 (3.1)	105.5 (5.0)	116.2 (20.4)	118.9 (17.4)	143.2 (8.8)
C	112.2 (16.8)	84.2 (22.3)	105.3 (14.0)	149.0 (13.4)	125.2 (6.0)	142.9 (19.8)
D	83.2 (12.5)	96.4 (12.1)	108.6 (12.1)	120.6 (15.9)	141.1 (20.6)	161.6 (12.8)

Table 4.7: Mean fundamental frequency (Hz) of each word in EL speech with manual F0 control averaged across two sentences for two male and two female speakers (values in parentheses are standard deviations).

Although there was no significant difference among the F0 values of the four sentence versions at the initial word position for both male and female speakers, similar characteristics as in normal speech were observed. For example, if we look at the two statement versions at the initial word position, version A (with sentence-initial focus) increases to a mean value that is 27.9 (±7.2) Hz and 33.4 (±28.8) Hz higher than version B for male and female speakers respectively. This pattern holds for two sentences and all four speakers. These results indicate that sentence-initial focus results in an increased F0

value on the stressed word, although the degree of increase was much smaller compared to the normal voice condition.

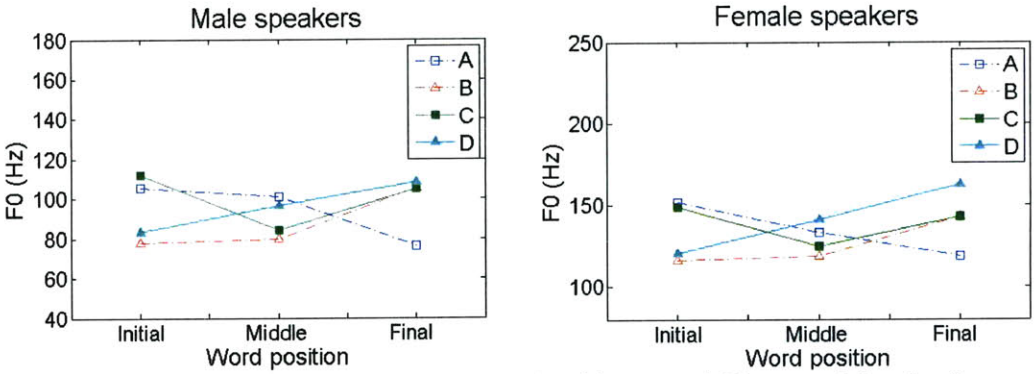


Figure 4.10: Average F0 peaks in EL speech with manual F0 control for the four sentence versions. Each peak represents the average for two sentences spoken by two male and female speakers. The sentence versions are described in Table 4.2.

The difference in the F0 patterns of the two question versions (C vs. D) was also observed at the initial word positions. Although the difference in F0 peak between versions was not significant, the frequency of version C was generally higher than that of version D. The exception was sentence 1 (Bev loves Bob) produced by speaker m2. At the final word location, version D was higher than version C only for three out of eight sentences, and the difference was not statistically different.

As a summary, similar characteristic of both word duration and F0 peaks as in normal speech was observed in manual F0 control condition. However, the effect of stress appeared to be less consistent and weaker in manual F0 condition compared to normal speech. A statistically significant difference in acoustic characteristics among the four different sentence versions was only observed for the F0 peak at the final word.

Relatively inconsistent use of acoustic cues in manual F0 condition would underlie the variability of performance in contrastive perception in this F0 condition.

More specifically, unlike the constant F0 condition, the dependency of performance on stress location was not consistent across speakers and sentences in Manual F0 control condition (Figure 4.11). For speaker f2, the performance for final stress was significantly higher compared to that for initial stress. If we further look at the result for this speaker for each sentence depending on the stress location as well as on the sentence type (Figure 4.12), it becomes clear that this is largely due to the very low percentage correct for in questions with initial stress (version C).

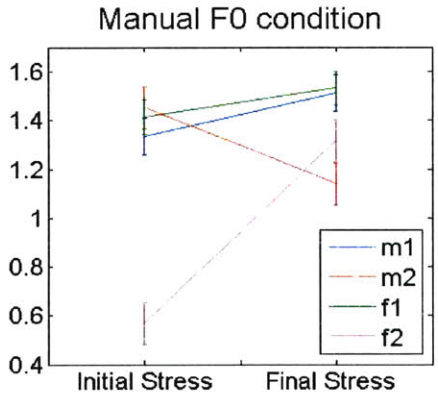


Figure 4.11: Mean and standard error of transformed percentage of correct responses of contrastive stress in EL speech with manual F0 condition.

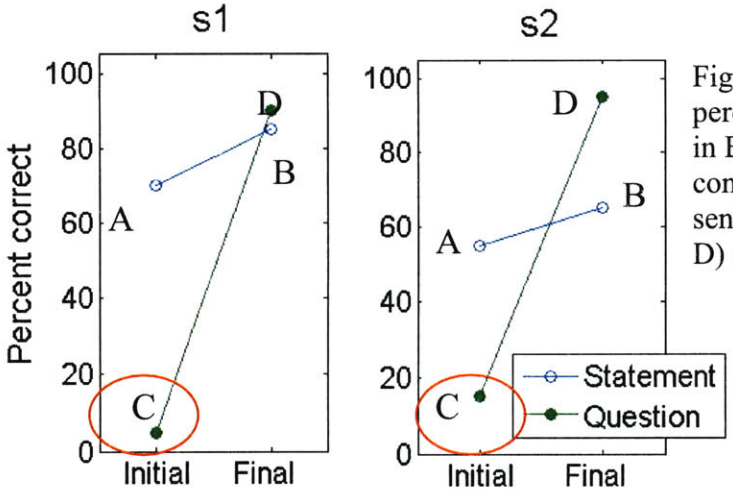


Figure 4.12: Percent correct for perception of contrastive stress in EL speech with manual F0 condition for speaker f2. The sentence versions (A, B, C, and D) are described in Table 4.2.

There seems to be durational cues for both sentence 1 and sentence 2 in manual F0 condition for this speaker (Figure 4.13) at the initial word position. The stressed word

was longer than the unstressed version for both sentences. On the other hand, unlike normal speech, there was no F0 peak at the initial word position for both sentences for version C (Figure 4.14), indicating that lack of F0 peak cue would be the major reason for the poor performance for s1 in question with initial stress for this speaker.

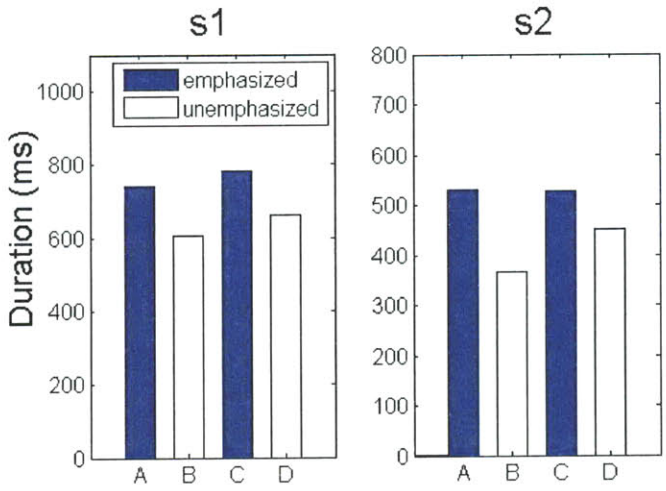


Figure 4.13: Word duration in EL speech with manual F0 control for the initial words for each sentence for speaker f2. The sentence versions are shown in Table 4.2.

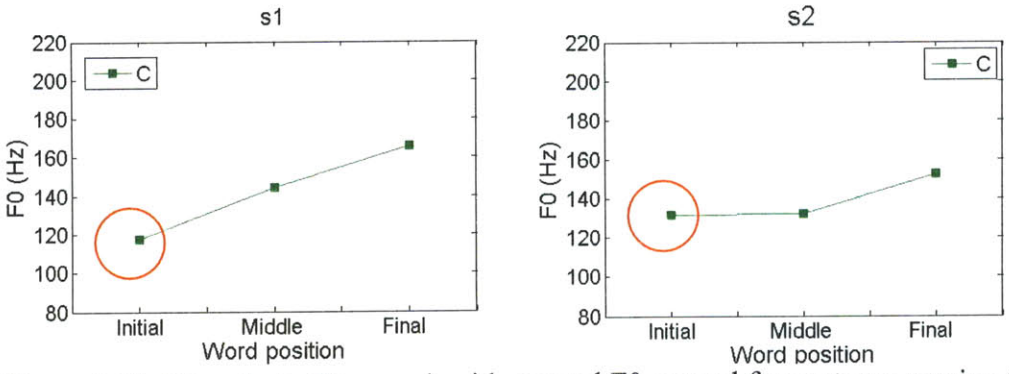


Figure 4.14: F0 peaks in EL speech with manual F0 control for sentence version C (question with initial stress) for each sentence for speaker f2.

As a second example, if we look at the performance for speaker m2, we can see that the performance at the final word position was lower compared to the performance at the initial word position. This characteristic was largely due to the low performance for sentence 1 in question with final stress (version D) (this sentence was perceived as having an initial stress although the speaker intended to have a final stress) (Figure 4.15). For

this case, durational cues (Figure 4.16) as well as F0 peak cues (Figure 4.17) were not well controlled. The stressed word was not longer than unstressed version, and the F0 peak was level throughout the utterance and not raised at the final stressed word.

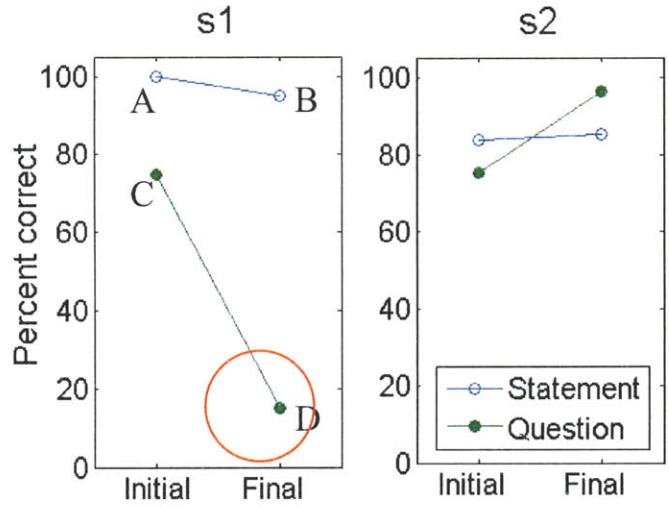


Figure 4.15: Percent correct for perception of contrastive stress in EL speech with manual F0 condition for speaker m2. The sentence versions (A, B, C, and D) are described in Table 4.2.

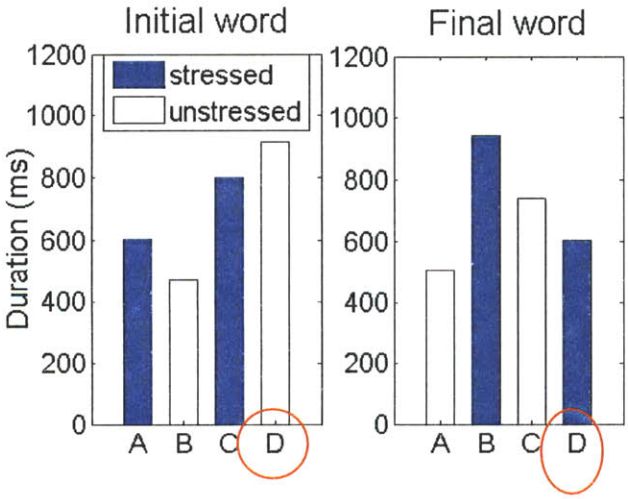


Figure 4.16: Word duration in EL speech with manual F0 control for the initial and final words for sentence 1 for speaker m2. The sentence versions are shown in Table 4.2.

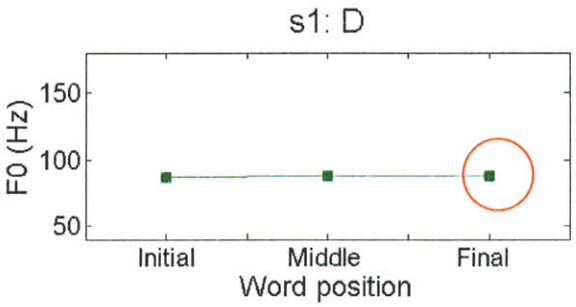


Figure 4.17: F0 peaks in EL speech with manual F0 control for sentence version D (question with final stress) for sentence 1 for speaker m2.

4.4.1.3 Amplitude-based F0 control condition

In amplitude-based F0 condition, the characteristics of word duration are the same as in EL speech with constant F0 condition described in section 4.4.1 (Figure 4.6). The results of the F0 peak analysis are presented in Table 4.8 and graphically represented in Figure 4.18. The results of statistical analysis did not indicate significant difference in F0 values among different sentence versions at initial and final words for both male and female speakers.

	Male			Female		
	First	Second	Third	First	Second	Third
A	84.8 (20.3)	94.7 (16.0)	94.8 (16.9)	142.8 (23.2)	153.7 (5.5)	159.7 (8.0)
B	91.6 (18.9)	103.6 (17.9)	107.6 (21.3)	136.8 (14.2)	155.8 (11.5)	162.7 (14.0)
C	89.7 (17.7)	97.5 (20.9)	94.7 (16.2)	140.5 (14.8)	150.7 (8.7)	150.1 (7.1)
D	87.3 (11.1)	89.4 (16.2)	100.8 (7.7)	133.8 (10.8)	144.0 (14.8)	157.0 (8.3)

Table 4.8: Mean fundamental frequency (Hz) of each word in EL speech with amplitude-based F0 averaged across two sentences for two male and two female speakers (values in parentheses are standard deviations).

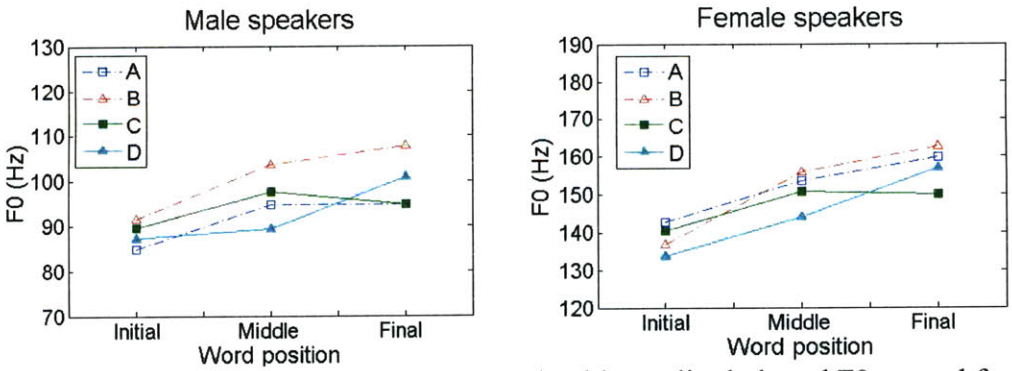


Figure 4.18: Average F0 peaks in EL speech with amplitude-based F0 control for the four sentence versions. Each peak represents the average for two sentences spoken by two male and female speakers. The sentence versions are described in Table 4.2.

Rather then prosodic effects, we found segmental effects on F0 peaks irrespective of the prosodic contexts. Figure 4.19 shows F0 peak values based on the amplitude for each initial word ('Bev', 'We') and final word ('Bob', 'away'). A separate ANOVA performed for each speaker showed significant differences in F0 peaks depending on the segmental contexts for speaker m1 (F'(3 ,12) = 11.56, MSE =690.26, p = 0.001) and speaker f1 (F'(3 ,12) =22.86, MSE =1849.16, $p < 0.001$) Pairwise comparisons (Bonferroni corrected) showed that, for speaker m1, F0 peak based on the amplitude for the vowel /i/ in 'we' was significantly lower than that for the vowel /a:/ in 'bob' and /eI/ in 'away' (p<0.01). For speaker f1, F0 peak based on the amplitude for /i/ in 'we' was significantly lower than that for /a:/ in 'bob', /ei/ in 'away', and /e/ in 'Bev' (p<0.01). The observed dependency on the segmental contexts resulted in higher F0 peaks for final words in this study and seemed to underlie the significantly higher performance for final stress (90.3%) compared to the performance for initial stress (47.5%).

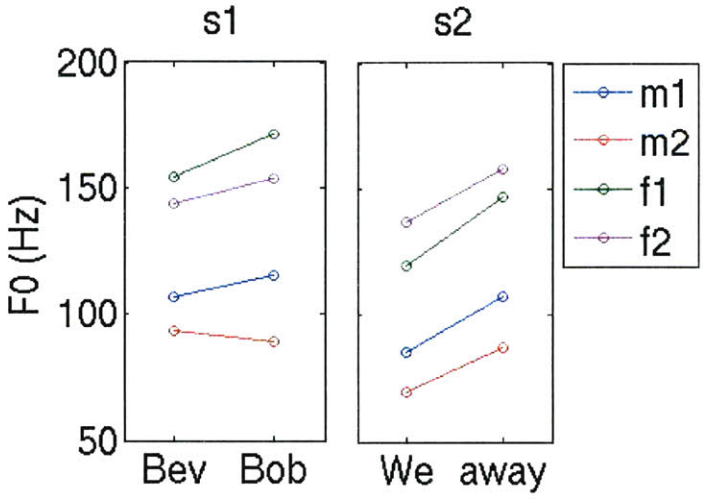


Figure 4.19: Average F0 peaks in EL speech with amplitude-based F0 control for each the four words. Each point represents the average for four sentence versions for each speaker.

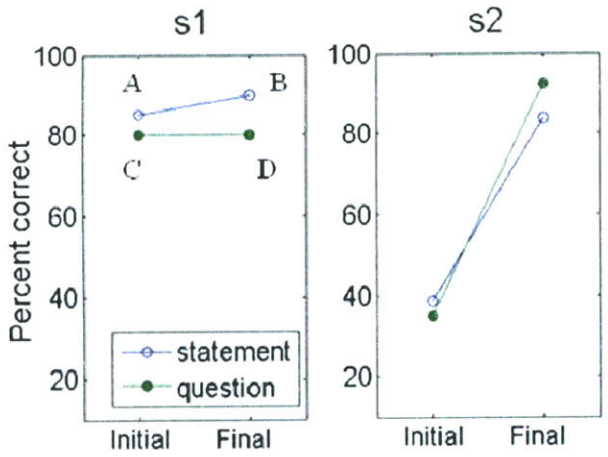


Figure 4.20: Percent correct for perception of contrastive stress in EL speech with amplitude-based F0 condition for speaker m2. The sentence versions (A, B, C, and D) are described in Table 4.2.

There was, however, a considerable variability across speakers and sentences. Figure 4.20 shows the performance of amplitude based F0 control for each sentence depending on the stress location and sentence type for speaker m2. While there was a large difference in performance between initial and final words for sentence 2 (s2), the performance for initial stress was comparable to final stress for sentence 1 (s1). Figure 4.21 shows the F0 peak for initial and final words for each version for each sentence. While in sentence 2 (s2), the F0 peak for final word based on the amplitude is higher than that for the final word irrespective of versions, it is not true for sentence 1 (s1). Especially the F0 peak for initial word was higher than that for final word in version C (question with initial stress). To investigate the influence of stress on the F0 peak, the data plotted in the left panel of Figure 4.21 was replotted for each stress location in Figure 4.22. This figure shows that the F0 peak was consistently higher in stressed words than in unstressed words for final words.

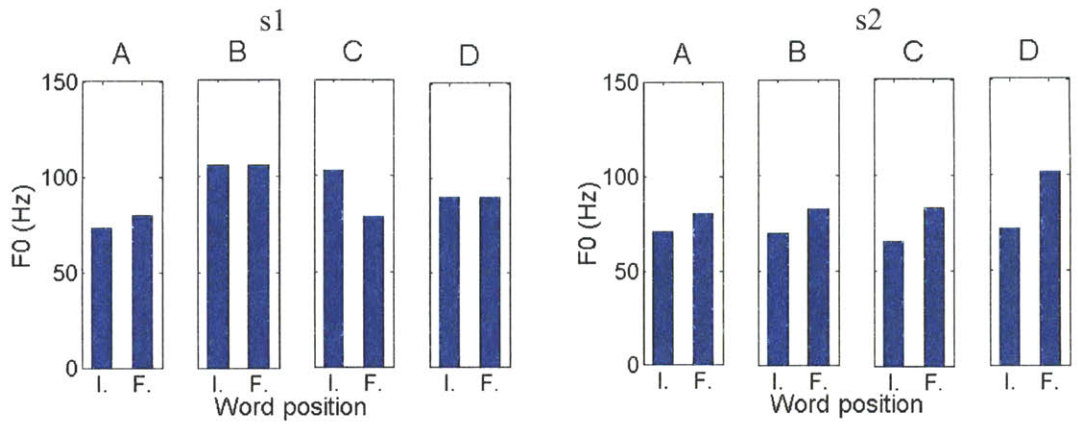


Figure 4.21: F0 peaks in EL speech with amplitude-based F0 control for the initial (I.) and final (F.) word positions for each sentence version. The sentence versions are described in Table 4.2.

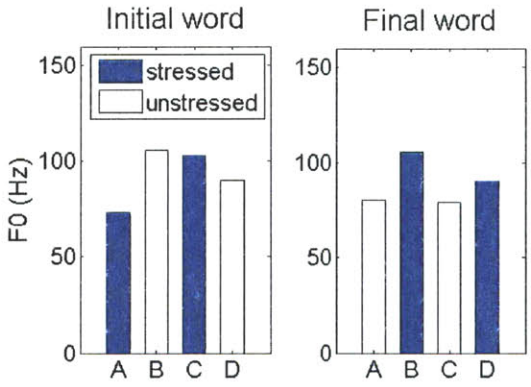


Figure 4.22: F0 peaks in EL speech with amplitude-based F0 control for the initial and final words for sentence 1 (s1) produced by speaker m2. The sentence versions are described in Table 4.2.

In order to examine whether this difference in F0 peak based on the amplitude is related to the supraglottal movement, first formant (F1) frequencies were measured at the vowel mid points in the final word 'Bob' in sentence 1 produced by speaker m2 and compared to those in the corresponding normal speech (Figure 4.23). Table 4.9 shows the average F1 measured at mid-vowels for each vowel in EL and normal speech conditions. The stressed low vowel /a:/ had higher F1 value compared to the unstressed condition in both EL and normal speech, implying that F1 frequencies and their corresponding articulatory configurations in terms of jaw, lips , and tongue positions might be manipulated to signal the linguistic structure of the prosody.

Furthermore, the degree of increase in F1 due to the contrastive stress seemed to be larger for EL speech compared to normal speech.

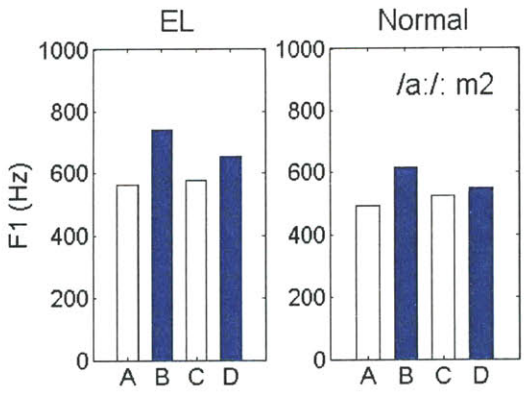


Figure 4.23: First formant (F1) frequencies measured at the mid-point in vowel /a:/ in 'Bob' in sentence 2 for EL speech with amplitude-based F0 control and in normal speech produced by speaker m2. The sentence versions are described in Table 4.2.

	m2	
	Stressed	Unstressed
EL speech	696.3	570.2
Normal speech	582.3	509.4

Table 4.9: Mean F1 frequency (Hz) measured at mid-point of vowel /a:/ in 'Bob' in sentence 1 in EL speech with amplitude-based F0 control and in normal speech for speaker m2.

4.4.2. Intonation

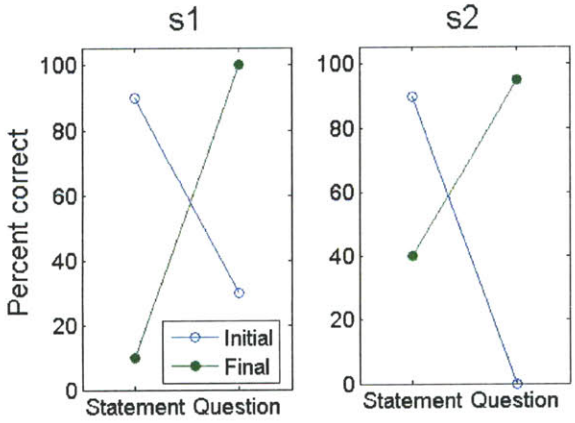


Figure 4.24: Percent correct for perception of intonation in EL speech with manual F0 control condition for speaker m1.

The fundamental objective in this section was to examine the underlying acoustic characteristics of the interaction in perception of intonation in manual F0 control

condition. The typically observed interaction is illustrated in Figure 4.24. The sentences with final stress tended to be perceived as questions, whereas the sentences with initial stress perceived as statements. In this section, the possible acoustic basis for this interaction is examined in EL speech with manual F0 control condition.

4.4.2.1. Question vs. statement in sentences with final stress.

In normal speech production, question intonation is signaled by a sharp F0 rise at the end of the utterance regardless of the position of the stress (Figure 4.25). However, as we discussed in Chapter one, due to the built-in constraints, the F0 cannot be ended at an arbitrary F0 value at the end in EL speech with manual F0 control. For example, the F0 contours produced by speaker m1 for the question with final stress version are shown in Figure 4.26. In these cases, unlike in normal speech, the F0 contour was not raised at the end of the final word, but rather the F0 was raised by approximately 60 Hz at the beginning of the final word and retained at that high F0 value for about 500 ms (471.9 ms for sentence 1 on the left and 499.7 ms for sentence 2 on the right) until it went down to the baseline F0 value just before the end of the final word. Listeners' accuracy in intonation perception for these sentences was 100 % and 95 % for sentence 1 and sentence 2, respectively indicating that different types of F0 contour shapes could still give a question intonation with a high degree of accuracy.

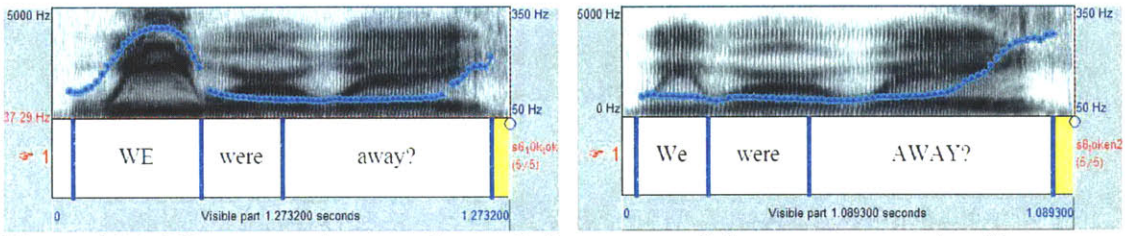


Figure 4.25: F0 contours in normal speech for sentence 2 for sentence version C (question with initial stress) on the right and for version D (question with final stress) produced by speaker m1.

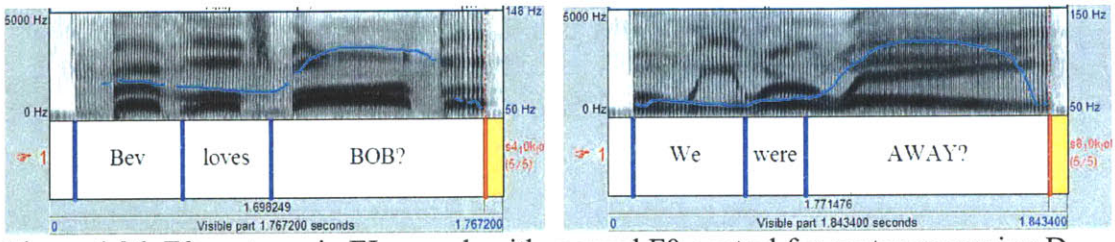


Figure 4.26: F0 contours in EL speech with manual F0 control for sentence version D (question with final stress) for sentence 1 (on the left) and sentence 2 (on the right) produced by speaker m2.

Similar F0 contours were observed for the statement with stress on the final word. The examples are shown in the Figure 4.27. In these cases, again the F0 was raised at the beginning of the final word by approximately 30 Hz and retained at the same level for 358.8 (s1, m1), 633.3 ms (s1, f2), 584.8 ms (s1, m2), and 550.0 ms (s2, m2) before going down at the end. The percentages of correct responses for these sentences were very low: 10% (s1, m1), 20% (s1, f2), 5% (s1, m2), and 0% (s2, m2), indicating that these F0 contours were perceived as questions rather as statements.

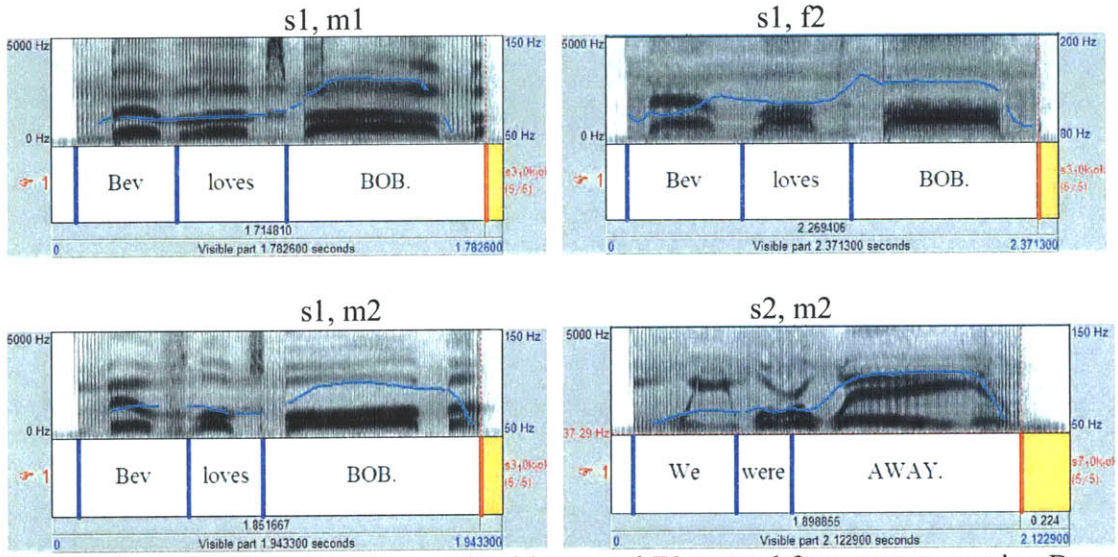


Figure 4.27: F0 contours in EL speech with manual F0 control for sentence version B (statement with final stress).

There was another type of F0 contour which again did not have a sharp F0 rise at the end but still produced a question intonation (Figure 4.28). In these two cases, the overall shape of the contours turned out be very similar to those obtained in the normal voices for statement with final stress (Figure 4.29). Despite the gross similarity in overall shape of these F0 contours, the rate of change in F0, measured from the beginning of the F0 rise to peak (marked by red lines) for 'away' did differ between EL speech with manual F0 control and normal speech. F0 rose more gradually on 'away' in EL speech with manual F0 than in normal speech. The summary of F0 differences between the beginning of F0 rise and F0 peak, and F0 slopes are shown for each condition and each sentence (Table 4.10). On inspecting this table, it becomes obvious that the F0 contour produced by the manual control is characterized by a quite gradual slope of F0 rise at the end.

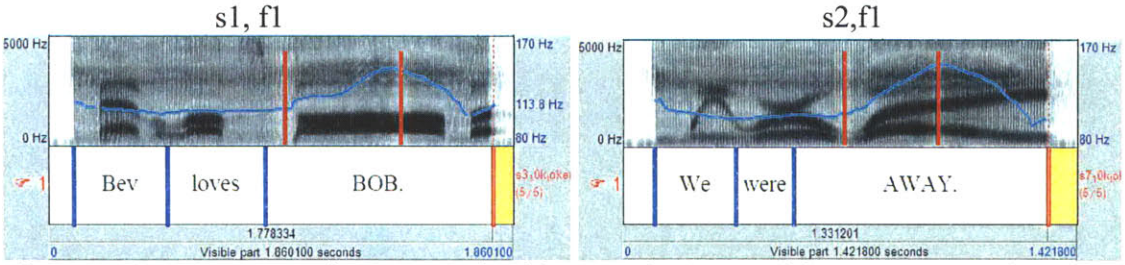


Figure 4.28: F0 contours in EL speech with manual F0 control for sentence version B (statement with final stress).

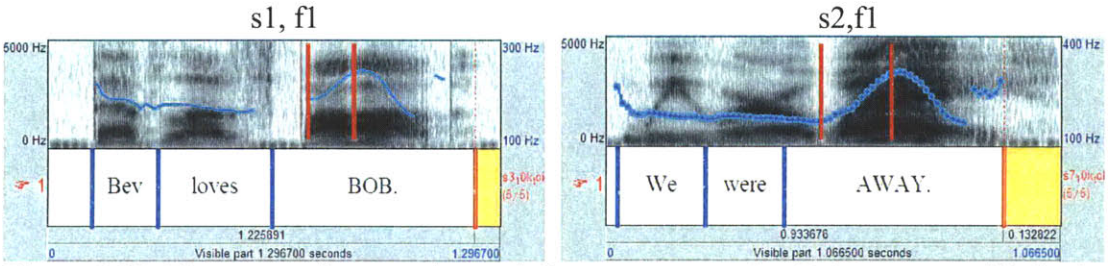


Figure 4.29: F0 contours in normal speech for sentence version B (statement with final stress).

		ΔF0 (st)	Δtime (s)	F0 slope
s1	Normal	4.2	0.15	28.4
	Manual	4.0	0.40	9.9
s2	Normal	10.2	0.17	60.8
	Manual	5.9	0.36	16.4

Table 4.10: Difference in F0 (semitone) and time as well as F0 slope between the beginning of the F0 rise and peak during the final word in normal speech and in EL speech with manual F0 control for speaker f1. F0 (semitone) was calculated by taking the F0 value at the beginning of the F0 rise as the base frequency.

4.4.2.2. Question vs. statement in sentences with initial stress.

Another important aspect of the intonation perception in manual F0 control is the significant degradation for the sentence version C (question with initial stress). Table 4.11 shows the performance for each speaker and sentence for this version. The examples which showed relatively low percentages in correct responses are shown in Figure 4.30.

	m1		m2		f1		f2	
	s1	s2	s1	s2	s1	s2	s1	s2
Intonation	30	0	50	15	30	55	95	95
stress	75	100	75	95	95	90	5	15

Table 4.11: Percent correct for perception of intonation and contrastive stress for sentence version C (Question with initial stress) for each speaker and sentence.

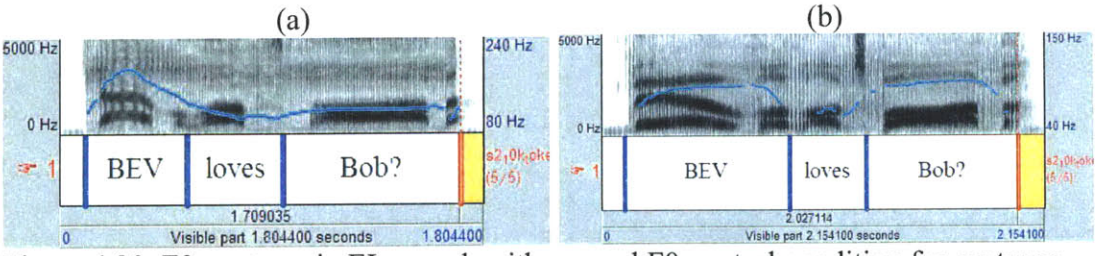


Figure 4.30: F0 contours in EL speech with manual F0 control condition for sentence version C (question with initial stress) which achieved relatively low performance: (a) F0 contour produced by speaker f1 with accuracy of 30%, (b) F0 contour produced by speaker m2 with an accuracy of 50%.

The degradation in performance seems to be either due to a) the insufficient increase in F0 at the end (four sentences) or due to (b) the F0 peak at the initial word (two sentences). In the latter case, we observed similar F0 contours at the final word as those observed for the sentences with final stress (Figure 4.27). The F0 was raised at the beginning of the final word by approximately 30 Hz and retained at the same level for 455.1 ms before going down to the baseline F0 frequency at the end. In this time, however, the percent correct was at a chance level (50%) probably due to the presence of F0 peak and durational increase at the initial stressed word.

The tokens which achieved a high level of accuracy (95%) (Figure 4.31) also showed a moderate rise in F0 from the second to final word. However, in this time, there was no F0 peak at the initial word, and the location of the stress was not correctly communicated. Table 4.11 shows the accuracy in perception of intonation and contrastive stress for each token. This table shows that those sentences which had a relatively lower percentage correct in the intonation task achieved a relatively higher performance in stress perception and vice versa, illustrating the difficulty associated with this type of sentence version (question with initial stress). For this type of intonation, it seemed to be possible to achieve a relatively high performance in either intonation or contrastive stress, but not in both prosodic contexts at the same time.

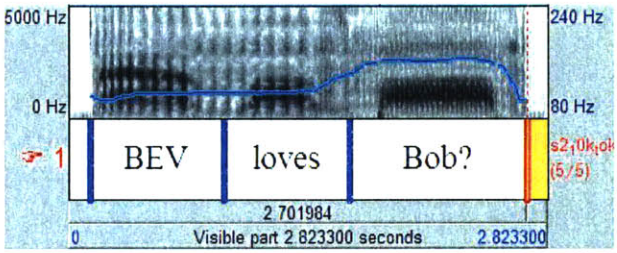


Figure 4.31: F0 contour in EL speech with manual F0 control condition for sentence version C (question with initial stress) which achieved relatively higher performance (90%) produced by speaker f2.

4.5. Discussion

In order to identify the possible acoustic basis for the performance observed in Chapter three, the speech rate, word duration, F0 peak, and formant frequencies were examined in three different EL speech conditions and normal speech produced by the same normal speakers. Results obtained for the contrastive stress will be examined first followed by the results for the intonation.

4.5.1. Acoustic characteristics and perception of contrastive stress

In constant F0 condition, the importance of durational cue for contrastive stress perception was confirmed as in the previous studies (Gandour & Weinberg, 1982). Although the speech rate in EL speech was significantly slower than in normal speech, the word duration was still found to be significantly longer when they were stressed compared to unstressed condition at the initial word position. However, the degree of durational increase due to stress was not necessarily larger in EL speech compared to normal speech. There seems to be thus no temporal exaggeration as a compensation for the missing F0 variation.

It was further shown that the reduction in perception of contrastive stress in final words compared to initial words seemed to be related to the differences in durational increase as a function of the word position. The smaller durational increase in stressed final words was probably partly due to the inherently longer segmental duration compared to initial words in the speech materials used in this study. As discussed by Cooper et al. (1984), characteristically longer words may undergo a smaller percentage

increase in duration. The smaller durational increase on the final word of the sentences may be also due to the effect of sentence-final lengthening. It has been shown that words at the end of a sentence have a longer duration than they do in other sentence positions (Klatt, 1975, 1976, Oller, 1973). Therefore, it may be more difficult to signal stress at the final word position by just the durational cue than at other word positions.

The analysis of the EL speech in amplitude-based F0 condition revealed inherent limitations associated with this control scheme as we expected. More specifically, the F0 peaks based on the amplitude of EL speech output was dependent on the segmental contexts rather than on the suprasegmental aspects: high vowels such as /i/ showed lower F0 peaks based on the amplitude compared to low vowels such as /a:/ in 'Bob'. The results of constant F0 and amplitude-based F0 together indicates that although durational cue alone can signal contrastive stress to some extent, F0 is the dominant cue and conflicting F0 evidence can override other cues. This result is consistent with the previous studies which examined listener perception of synthetic speech (Denes, 1959; Denes & Milton-Williams, 1962).

However, examination of the vowels /a:/ in "Bob" produced by one male speaker revealed that F0 calculated based on the amplitude was consistently higher for stressed words compared to unstressed words. We also found some difference in F1 frequencies depending on whether the word was stressed or not. This effect appeared to be more pronounced in EL speech than in normal speech in the speaker examined. The resulting formant frequencies could be understood to some extent as a consequence of the effect that changes in jaw and tongue position would have on the shape of the vocal tract. The jaw and lip opening maxima are likely to represent the degree of the vocal tract opening

that is related to sonority expansion, whereas the tongue movement together with the acoustic data can indicate how place features are phonetically realized (Kent & Netsell, 1971). Since no articulatory data (jaw and tongue dorsum) were measured in this current study, it is not possible to address the question whether the formant movement reflects the movement of the tongue dorsum, lips, or jaw. Even if we find significant differences for /a/, there is still a question whether greater jaw opening due to hyperarticulation would be useful for amplitude-based F0 modulation. In particular, although hyperarticulation would yield more open /a/, it should yield more constricted /i/, and the effect might depend on the exact vowel quality. If, on the other hand, the effects are solely due to greater duration, and duration tends to make all vowels more open, then this could provide a positive support for the amplitude-based F0 modulation. To further explore the prosodic effects on vowel articulation we need to examine the acoustic and articulatory characteristics of different vowel types spoken by a larger number of speakers.

In manual F0 condition, the analysis of acoustic results and perceptual performance showed that most of the selected tokens were sufficient to convey contrastive stress. This was achieved by manipulating the appropriate cues in a manner resembling normal speech. In those instances when EL speakers with manual F0 control produced the opposite effect of what was expected, such as shorter word duration and lower F0 peak on the stressed word, listeners' responses were no longer accurate.

We also observed that the degree of lengthening of word duration due to the presence of stress was less manifested and realized less consistently in manual F0 control condition compared to normal speech and EL speech with constant F0 condition. This

would be probably because the speakers were focusing on manipulating F0 contour with manual control and were less attentive to the durational cue. Even though the use of durational cue appeared to be less consistent compared to those in normal speech and other EL speech conditions, manually controlled F0 peaks values seemed to be sufficient to produce high level of accuracy in perceiving location of contrastive stress, further supporting the dominant role of F0 cue in contrastive stress perception.

4.5.2. Acoustic characteristics and perception of intonation

In the manual F0 condition, when it was attempted to produce a stress on the final word, the sentences tended to be perceived as questions rather than statements irrespective of the speaker's intentions and these perceptual performance were realized by different types of F0 contours compared to the F0 contours generally seen in the normal speech. The result indicates that the F0 contour does not have to be raised at the end as in the normal speech, but either needs to be retained at the higher F0 value for a while or needs to be raised with a slower rate. Thus, although the phonetic realization of prosodic structure in EL speakers was dissimilar when compared to the phonetic realization of prosodic structure in normal speakers, listeners still perceived the intended prosodic structure in EL speakers as accurately as the intended prosodic structure in normal speakers. This finding supports the view that listeners identify an intonation pattern by encoding not only endpoint F0, but also various aspects of F0 contours, such as rate of change in F0, previous F0 peak, and the overall shape of the contour (Hadding-Koch & Studdert-Kennedy, 1964; Majewski & Blasde, 1969).

Chapter 5

Summary and Discussion

5.1 Summary of the Findings

In this dissertation, an approach for amplitude-based control of F0 in EL speech was developed and its impact on the quality of EL speech was first examined. We further examined the prosodic control abilities of amplitude-based F0 control to convey the distinction between question and statement intonation and location of contrastive stress using a perceptual identification task. The results in perceptual experiments were compared with those in EL speech manual F0 control and the underlying acoustical characteristics were investigated.

The results of perceptual experiments showed a significant improvement of the overall naturalness of declarative sentences, but also revealed inherent limitations in communicating linguistic contrasts in amplitude-based F0 control scheme. The F0 based on the amplitude of EL speech wave appeared to depend on segmental contexts rather than on the suprasegmental aspects. On the other hand, the performance of selected tokens in manual F0 control showed generally a high performance in communicating contrastive stress. The problem with the manual F0 control was the great variability across repetitions and speakers. The results of the acoustical analysis of the EL speech showed that although in general, stressed word was lengthened in EL speech as well as

normal speech, the percentage increase in duration was smaller and the effect of contrastive stress on the duration was less consistent in EL speech, especially with manual F0 control condition. Similar trends were observed for the analysis of F0 contours in EL speech with manual F0 control. Although the stressed word generally had higher F0 peaks than did unstressed word, the difference in F0 was relatively smaller and less consistent compared to normal speech and seemed to underlie the types of errors in perception of contrastive stress in EL speech with manual F0 control.

We also noted that the performance for intonational distinction depended on the particular prosodic context in EL speech with manual F0 control. For the question with initial stress and the statement with final stress, speakers could communicate either stress location or intonation relatively easily. However it became drastically difficult if they had to communicate both stress location and intonation at the same time. The results of the acoustical analysis together with the perceptual characteristic for intonation indicate that the difficulty associated with communicating intonation in EL speech with manual F0 control may not necessarily due to the built-in limitation that F0 cannot be ended or started at an arbitrary F0 value. Even though there was not a sharp F0 rise at the end of the utterances as in normal speech, speakers were still able to communicate the question intonation with a high percentage correct. The problem seems to be rather more complicated due to the significantly slower speech rate and may have to be analyzed in terms of the difficulty in coordinating a particular F0 shape with the temporal structure of utterances so that the distinction between statement and question can be effectively communicated.

6.2. Limitations of the Current Study

Although the investigation in Chapter two demonstrated preliminary feasibility of the amplitude based F0 control of an EL, it was meant to essentially demonstrate a proof-of-concept, and was therefore limited with respect to number of subjects, sentences, and stimuli used in the perceptual experiments. Thus, the generalisability of the results must be viewed with caution. We also did not test whether just any variation in EL F0 that was not linked to amplitude would also produce a similar level of preference. For example, just any random F0 fluctuation may improve the naturalness of the EL speech compared to the constant F0 condition. It is also possible that a simple addition of declination may help improve the perceived naturalness of the EL speech, but may not help with linguistic contrasts. More research is needed to verify these possibilities.

In the investigation in Chapter three and four, the stimulus material was chosen so that individual phrases were short and only voiced consonants were included. Thus, possible limitations in the EL speech were accommodated as much as possible. Considering the relatively poor performance in rather complicated prosodic structure for intonation in EL speech with manual F0 control condition (question with initial stress and statement with final stress), increasing the target-sentence length and including more demanding stimulus material in terms of phonetic context of target sentences, grammatical categories of words in the sentences, and number of contrastive stress categories, may introduce more difficulty marking the contrast.

6.3. Future Perspectives

As the results of the VAS revealed in Chapter two, the rating for the best token, EL_f0n (EL speech with F0 modulation based on the normal F0 contour), was to the right of the mid-point of the scale (towards the "Very Different" end), suggesting that there were still other important acoustic factors that need to be addressed to improve the quality of the EL speech in addition to F0 modulation. This finding is consistent with the previous studies on the enhancement of the EL speech (Melztner, 2003; Meltzner & Hillman, 2005). Other important acoustic properties include deficits due to the acoustic characteristics of the EL voicing source and its location away from the terminal end of the vocal tract (i.e., introduction of spectral zeroes into the speech output), and additional modifications in the vocal tract transfer function due to the impact of the laryngectomy operation on the upper airway (Meltzner, 2003; Myrick & Yantorno, 1993). The analysis-by-synthesis approach developed in chapter 2 (details are in appendix 1) using KLSYN should provide the means for investigating (via generating stimuli for perceptual experiments) and testing attempts to correct (via modifying synthesis parameters) additional acoustic deficits in EL speech.

Another area of inquiry has to do with the potential source of amplitude fluctuation in EL speech. One potential source of the amplitude variation in the EL speech is changes in mouth opening during speech production, but this did not always seem to account for the magnitude of the observed variation. It is possible that the user manipulates the pressure of the EL device against the neck, in a manner similar to a body or hand gesture that occurs during speech production. This manipulation could influence the pressure against the neck and therefore modify the amplitude of the

acoustic source that excites the vocal tract. Rothman (1982) have initiated a series of experiments that were designed to examine the acoustic characteristics of the various artificial larynges and the effects of their coupling to the neck. It was noted that there was an increase in the intensity of the energy between 3000-6000 Hz when coupling pressure was increased. Another possibility is that the low frequency deficit of the EL device decreases the first formant amplitude of high vowels more than low vowels, so there is a vowel-dependent fluctuation in amplitude.

Additional studies of pre-recorded EL speech (in digital audio and video format) from patients with laryngectomies (Goldstein, 2003) can be conducted to evaluate hypothesized changes in amplitude due to movements of formant frequencies, changes in formant bandwidths, the degree of low frequency deficit, and the degree of mouth opening. To examine the potential role of EL location and contact pressure, new recordings of laryngectomy EL users can be made using video recordings and a sensor on the head of the EL to measure the pressure exerted against the neck so that changes in coupling pressure can be quantified and correlated with acoustic output. A clearer understanding of the sources could potentially lead to improved algorithms for real-time enhancement of EL speech based on processing of the EL speech output. It could also suggest ways of training an EL user to manipulate the device to produce more natural prosody.

Another possible future addition to this work is to examine the implications for the laryngectomized patients who are native speakers of tone languages. As reviewed in the introduction, a lack of adequate F0 control has largely limited the ability of the EL users to signal tonal contrasts (Gandour et al., 1988; Liu et al., 2006; Ng et al., 2001). In

this context, it is interesting to note that in Mandarin Chinese, amplitude has been suggested to contribute to tone recognition, when F0 information was removed (Fu & Zeng, 2000; Liu & Samuel, 2004; Whalen & Xu, 1992). It has been further demonstrated that this amplitude-based tone recognition was directly related to the correlation between amplitude contour and F0 contour, indicating that subjects might have interpreted amplitude changes as F0 changed (Fu & Zeng, 2000). More research on the acoustic characteristics of tone languages in EL speech might be needed to extend the scope of our study for tone languages.

It would be also interesting to examine the rate-related changes in the acoustic-phonetic structure in EL speech. Although we found significantly slower speech rate in EL speech with constant F0 and manual F0 control condition than in normal speech, consonant and vowel durations in an utterance may not be increased or decreased uniformly (Miller, 1981). It should be also noted that, in the acoustic analysis described in Chapter four, the F0 contours were mostly characterized by the F0 peak values in each word. It has been documented that the perception of intonation is also affected by where the F0 peak is located (Cooper et al., 1985). More specifically, the similar F0 contour (rise-fall, for example) can be perceived as either statement or question depending on where the F0 peak is located in the word. This aspect was not quantitatively examined in this study and should be investigated in the future study to more fully examine the effects of different F0 control strategies on the ability to communicate the linguistic contrast.

Considering the limitation of amplitude based F0 cue to communicate linguistic contrast, another important future work may include exploring other acoustic cues which might be correlated with speakers' efforts in communicating linguistic contrasts. In

normal speech, it has been found that not only suprasegmental features (duration, pitch, loudness), but also segmental phonetic cues are manipulated to signal the linguistic structure of prosody. More specifically, stressed vs. unstressed vowels are differentiated both by their formant patterns and their articulatory configurations in terms of jaw and tongue dorsum positions. The jaw moves lower and the tongue moves more in the direction of the phonological specifications of the vowel. In EL speech, it may be also possible that speakers try to manipulate not only the temporal cues but also articulatory movements in order to communicate the prosodic contrast. Since there is no F0 cues available in EL speech with constant F0, the degree of strengthening can be even more exaggerated compared to normal speech.

It has been also shown that intonation is not only related to the F0 pattern of the speech wave, but also to changes along other acoustic dimensions (Denes, 1959; Higashikawa, Nakai, Sakakura, & Takahashi, 1996; Higashikawa, & Minifie, 1999; Meyer-Eppler, 1957). Meyer-Eppler (1957), for example, showed that in whispered speech the third formant was affected by the intonation. Denes (1959) also showed that the movement of the third formant was related to changes of intonation pattern in normal speech and found that the changes were much more pronounced with whisper than with voiced speech. The relationship may deserve further investigation in EL speech.

The investigation discussed in this document was conducted with the aim of using the results to guide a future enhancement effort to improve the quality of EL speech. Despite the limitation in terms of communicating linguistic contrast, considering the significant improvement in overall naturalness and the fact that the mechanical quality of the EL speech has been the one of the primary concerns of the EL users, it would be

worth investigating further the implementation of the amplitude-based F0 control scheme in real-time and examining its impact on the quality of life of laryngectomy patients. One possible configuration could entail using the DSP system to estimate the RMS amplitude of EL speech from a microphone signal, and to then generate an F0 contour (based on linear prediction) that could be fed back to drive the EL device in a real-time loop.

Appendix

Synthesizing EL speech using the Klatt formant synthesizer

The procedure used as a first pass to synthesize EL speech using the Klatt synthesizer was based on (Hanson, 1995; Klatt, 1980; Klatt, Chapter 3; Klatt & Klatt, 1990). The synthesis was performed using the KLGlott88 voicing source. The synthesis sampling rate was set to 11.5 kHz for speaker 1 and 12 kHz for speaker 2 based on average formant spacing, which was determined by preliminary spectral analysis. The laryngectomy patients' vocal tracts have been truncated and thus the formant frequencies tend to be shifted higher in frequency (Sisty & Weinberg, 1972). Average formant spacing for the EL speech tended to be greater than 1000 Hz (Meltzner, 2003) and in order to include five formants of EL speech in the synthesis, a synthesis sampling rate greater than 10 kHz was necessary to avoid aliasing effects (Hanson, 1995; Klatt, Chapter 3). The original sampling rate was 32 kHz and, all speech samples, including the pre-laryngectomy speech, were lowpass filtered and downsampled to a 12 kHz sampling rate for speaker 1 and 11.5kHz sampling rate for speaker 2 using the software program Wavesurfer, and were brought into Xkl software tool for analysis, synthesis, and modification.

Most synthesis parameter estimates were derived from matching the time-varying spectra of the original utterances. All spectral measurements were performed using the Xkl software tool. The frequency measurements were taken from individual DFT spectra with a 40 ms Hamming window. Waveform displays were used to determine durations and locations of speech segments. Spectrograms were also used to visualize the

comparison of the original and synthesized speech. The segmental durations were specified by hand in the synthesis so as to match the original recording.

As a preliminary step, local spectra were obtained at mid-vowel, vowel onsets, and offsets. Short segments of speech were synthesized to match these local spectra, with all parameters set to be constants. Spectra of this initial copy synthesis were obtained to begin an iterative process of measuring the spectral differences between the original and synthesized speech, then adjusting synthesis parameters to improve the spectral matching. The parameters so determined were later used to synthesize complete vowel segments with time-varying parameters. Open quotient (OQ), defined as the percentage of time the glottis was open in one fundamental period, was adjusted to be very low (~12 %) to mimic the low frequency deficit of EL speech. Spectral tilt (TL) and formant bandwidths were adjusted to match the relative formant amplitudes. To introduce the zeros associated with the placement of the EL device, the nasal and/or tracheal pole-zero pairs available in the synthesizer were placed at the measured zero frequencies. The bandwidths of the poles and zeros were adjusted such that the bandwidths of the zeros were narrower than those of the poles. Finally the amplitude of voicing (AV) was adjusted to match the overall energy of the synthetic speech to the original. Examples of spectra are shown in Figure A1. Compared to the parameters used for the natural speech (Klatt & Klatt, 1990), EL speech involved smaller OQ, lower overall TL, and fixed F0. A nasal pole-zero pair was also introduced in order to achieve the detailed low frequency characteristics of the EL speech. EL speech also required relatively higher formant frequencies and narrower bandwidths.

Once a reasonable match to the original spectra was obtained, synthesis parameters were then varied as a function of time, including the first five formant frequencies (F1, F2, F3, F4, F5), their corresponding bandwidths (B1, B2, B3, B4, B5), and four source parameters: F0, AV, TL, and OQ. The amplitude variation in the output speech was realized by varying the amplitude of voicing (AV). The input data were reduced by selecting values at specific points in time for each parameter so that the linear interpolation generated by the synthesizer between adjacent points would capture the kinematics as closely as possible. Using this approach, it was possible to closely match the acoustic characteristics of EL speech with synthesized replicas. Informal listening also provided subjective confirmation of the similarity of the original and synthesized speech.

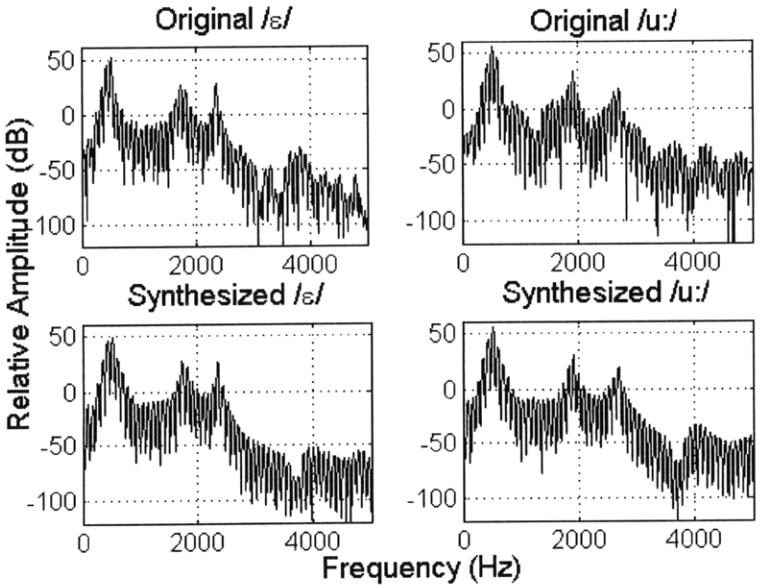


Figure A1: Comparison of spectra obtained at mid-vowel in EL speech. The original /ɛ/ (top left) vs. synthesized /ɛ/ in "Mary" (bottom left). The original /u:/ (top right) vs. synthesized /u:/ (bottom right) in "too" (sentence 1, speaker 1).

Bibliography

Atkinson, J.E. (1973). Aspects of intonation in speech: Implications from an experimental study of fundamental frequency, unpublished Ph.D. thesis, University of Conneticut.

Atkinson, J. E. (1976). Inter- and intraspeaker variability in fundamental voice frequency. *Journal of Acoustical Society of America, 60,* 440-445.

Bangayan, P., Christopher, L., Alwan, A. A., Kreiman, J., & Gerratt, B. R. (1997). Analysis by synthesis of pathological voices using the Klatt synthesizer. *Speech Communication, 22,* 343-368.

Bennett, S., & Weinberg B. (1973). Acceptability ratings of normal, esophageal, and artificial larynx speech. *Journal of Speech, Language, and Hearing Research,*16(4): p. 608-615.

Boersma, P., & Weenik, D. (2005). *Praat, a system for doing phonetics by computer, Version 5.1.10.* Amsterdam: University of Amsterdam, Institute of Phonetic Sciences. Available online at http://www.praat.org.

Choi, H. S., Park, Y. J., Lee, S. M., & Kim, K. M. (2001). Functional characteristics of a new electrolarynx "Evada" having a force sensing resistor sensor. *Journal of Voice, 15,* 592-599.

Cole, D., Stridharan, S., Moody, M., & Geva, S. (1997). Application of noise reduction techniques for alaryngeal speech enhancement. *Proceedings of IEEE TENCON '97. IEEE Region 19 Annual Conference Speech and Image Technologies for Computing and Telecommunications, 2,* 491-494.

Copper W. E., Eady S. J., & Mueller P. R. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of Acoustical Society of America, 77,* 2142-2156.

Denes, P. (1959). A preliminary investigation of certain aspects of intonation. *Language and Speech, 2,* 106-122.

Denes, P., & Milton-Williams, J. (1962). Further studies in intonation. *Language and Speech, 5,* 1-14.

Delattre, P (1969). An acoustic and articulatory study of vowel reduction in four languages, International Review of Applied Linguistics. 7, 295-325.

Eady, D. J., & Cooper, W. E. (1986). Speech intonation and focus location in matched statements and questions. *Journal of Acoustical Society of America, 80,* 402-415.

Espy-Wilson, C. Y., Chari, V. R., MacAuslan, J. M., Huang, C. B., & Walsh, M.J.

(1998). Enhancement of electrolaryngeal speech by adaptive filtering. *Journal of Speech, Language, and Hearing Research, 4*, 1253-1264.

Fant, C. G. M (1970). Acoustic theory of speech production. The Hague: Mouton.

Fu, Q-J., & Zeng, F.G. (2000). Identification of temporal envelope cues in Chinese tone recognition. *Asia Pacific Journal of Speech, Language, and Hearing, 5*, 45-57.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech, 1(2)*, 126-152.

Galyas, K., Branderud, P., & McAllister, R. (1982). The "intonator". Development of an electrolarynx with intonation control. In A. Sekey (Ed.), *Electroacoustic Analysis and Enhancement of Alaryngeal Speech* (pp.184-189), Springfield: Charles C Thomas Pub Ltd.

Gandour, J., & Weinberg, B. (1982). Perception of contrastive stress in alaryngeal speech. *Journal of phonetics, 10*, 347-359.

Gandour, J., Weinberg, B., & Kosowsky, A. (1982). Perception and syntactic stress in alaryngeal speech. *Language and Speech, 25*, 299-304.

Gandour, J., & Weinberg, B. (1983). Perception of intonational contrasts in alaryngeal speech. *Journal of Speech and Hearing Research, 26*, 142-148.

Gandour, J., Weinberg, B., & Garzione, B. (1983). Perception of lexical stress in alaryngeal speech. *Journal of Speech and Hearing Research, 26*, 418-424.

Gandour, J., & Weinberg, B. (1984). Production of intonation and contrastive stress in electrolaryngeal speech. *Journal of Speech and Hearing Research, 27*, 605 612.

Gandour, J., & Weinberg, B. (1986). Production of syntactic stress in alaryngeal speech. *Language and speech, 28*, 295-306.

Gandour, J., Weinberg, B., & Petty, S. H. (1986). Production of lexical stress in alaryngeal speech. *Folia Phoniatrica, 37*, 279-286.

Gandour, J., Weinberg, B., Petty, S. H., & Dardarananda, R. (1988). Tone in Thai alaryngeal speech. *Journal of Speech and Hearing Disorders, 53*, 23-29.

Goldstein, E. A. (2003). *Prosthetic voice controlled by muscle electromyographic signals.* Ph.D. Thesis, Harvard University, Cambridge.

Goldstein, E. A., Heaton, J. T., Kobler, J. B., Stanley, G. B., & Hillman, R. E. (2004).

Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity. *IEEE Transactions on Biomedical Engineering, 51*, 325-332.

Gray, S., & Konrad, H. R. (1976). Laryngectomy: postsurgical rehabilitation of communication. *Archives of Physical Medicine and Rehabilitation, 57*, 140-142.

Hadding-Koch, K., & Studdert-Kennedy, M. (1964). An experimental study of some intonation contours. *Phonetica, 11,* 175-185.

Hanson, H. (1995). Synthesis of female speech using the Klatt formant synthesizer. *MIT Speech Communication Group Working papers, 10*, 84-103.

Heaton, J. T, Goldstein, E. A., Kobler, J. B., Zeitels, S., Randolph, G., Walsh, M., et al. (2004).Surface electromyographic activity in total laryngectomees following laryngeal nerve transfer to neck strap muscles: Correlation with vocal and non-vocal behaviors. *Annals of Otology, Rhinology and Laryngology, 109*, 972-980.

Higashikawa, M. Nakai, K. Sakakura, A, & Takahashi, H. (1996). Perceived pitch of whispered vowels – relationship with formant frequencies: a preliminary study. *Journal of voice, 2*, 155-158.

Higashikawa, M. & Minifie, F. D. (1999). Acoustical-perceptual correlates of "Whisper Pitch" in synthetically generated vowels. *Journal of Speech, Language, and Hearing Research, 42*, 583-590.

Hillman, R. E., Walsh, M. J., Wolf, G. T., Fisher, S. G., & Hong, W. K. (1998). Functional outcomes following treatment for advanced laryngeal cancer. Part I—Voice preservation in advanced laryngeal cancer. Part II—Laryngectomy rehabilitation: the state of the art in the VA System. Research Speech-Language Pathologists. Department of Veterans Affairs Laryngeal Cancer Study Group. *Annals of Otology, Rhinology and Laryngology Supplement, 172*, 1-27.

Kakita, Y., & Hirama, J. (1989). Controls of prosodic information and voiceless consonants for the electronic larynx. *Technical report of IECE. SP88-148*, 25-30.

Kent, R. D., & Netsell, R. (1971). Effects of stress contrasts on certain articulatory parameters. Phonetica, 24, 23-44.

Kent, R. D., & Read C. (2001). *Acoustic analysis of speech.* Singular.

Kikuchi, Y., & Kasuya H. (2004). Development and evaluation of pitch adjustable electrolarynx. *Speech Prosody 2004*, 761-764.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in connected discourse. *Journal of phonetics, 3*, 129-140.

Klatt. D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of Acoustical Society of America, 59,* 1208-1221.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of Acoustical Society of America, 67,* 971-995.

Klatt, D. H. Description of the cascade/parallel formant synthesizer. Chapter 3 of book in preparation.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of Acoustical Society of America, 87,* 820-857.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36,* 21-40.

Lehiste, I. (1976). Suprasegmental features of speech. N. Lass (Ed.), *Contemporary issues in experimental phonetics (pp.*225-239). New York: Academic Press.

Liberman, M., & Peirrehumbert, J. (1984). "Intonational invariance under changes in pitch range and length," in *Language Sound Structure*, edited by M. Aronoff and R. T. Oeherle (MIT, Cambridge, MA).

Liu, H., Zhao, Q., Wan, M., & Wang, S. (2006). Application of spectral subtraction method on enhancement of electrolarynx speech. *Journal of Acoustical Society of America, 120,* 398-406.

Liu, S., & Samuel, A. G. (2004). Perception of Mandarin lexical tones when F0 information is neutralized. *Language and Speech, 47,* 109-138.

Liu, H., Wan, M., Ng. L. M., Wang, S., & Lu, C. (2006). Tonal perceptions in normal laryngeal, esophageal, and electrolaryngeal speech of Mandarin. *Folia Phoniatrica et Logopedica,* 58, *340-352.*

Ma, K., Demirel, P., Espy-Wilson, C., & MacAuslan, J. (1999). Improvement of Electrolarynx speech by introducing normal excitation information. *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH),* Budapest, 323-326.

Majewski, W., & Blasde, U, R. (1969). Influence of fundamental frequency cues on the perception of some synthetic intonation contours. *Journal ofthe Acoustical Society of America, 45,* 450-457.

Markel, J. D., & Gray, A. H. (1976). *Linear prediction of speech.* New York, NY:

Springer-Verlag. Mendenhall, W. M., Morris, C. G., Stringer, S. P., Amdur, R. J., Hinerman, R. W., Villaret, D. B., & Robbins, K. T. (2002). Voice rehabilitation after total laryngectomy and postoperative radiation therapy. *Journal of Clinical Oncology, 20*, 2500-2505.

McHenry, M., Reich, A., & Minifie, F. (1982). Acoustical characteristics of intended syllabic stress in excellent esophageal speakers. *Journal of Speech and Hearing Research, 25*, 564-573.

Meltzner, G. S (2003). *Perceptual and acoustic impacts of aberrant properties of 0 electrolaryngeal speech*. Doctoral dissertation, MIT, Cambridge.

Meltzner, G. S., Kobler, J. B., & Hillman R. E.(2003). Measuring the neck frequency response function of laryngectomy patients: Implications for the design of Electrolarynx devices. *Journal of Acoustical Society of America, 114*, 1035-1047.

Meltzner, G. S., & Hillman, R. E. (2005). Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech. *Journal of Speech, Language, and Hearing Research, 48*, 766-779.

Meyer-Eppler, W. (1959). Realization of prosodic features in whispered speech. *Journal of Acoustical Society of America, 29,* 104-106.

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In *Perspectives on the Study of Speech*, edited by P.D. Eimas and J.L.Miller (Erlbaum, Hillsdale, NJ).

Morris, H. L., Smith, A. E., Van Demark, D. R., & Maves, M. D. (1992). Communication status following laryngectomy: the Iowa experience 1984- 1987. *Annals of Otology, Rhinology and Laryngology, 101*, 503-510.

Morton, J., & Jassem, W. (1965). Acoustic correlates of stress. *Language and Speech, 8*, 159-181.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 9*, 453-467.

Myrick, R., & Yantorno, R. (1993). Vocal tract modeling as related to the use of an artificial larynx. *Bioengineering Conference Proceedings of the 1993 IEEE Nineteenth Annual Northeast*, 75-77.

Ng, M. L., Gilbert, H. R., & Lerman, J. W. (2001). Fundamental frequency, intensity, and vowel duration characteristics related to perception of Cantonese alaryngeal speech. *Folia Phoniatrics et Logopaedica, 53*, 36-47.

Niu, H. J., Wan, M. X., Wang, S. P., & Liu, H. J. (2003). Enhancement of Electrolarynx

speech using adaptive noise cancelling based on independent component analysis. Medical and Biological Engineering & Computing, 41, 670-678.

Okobi, A. O. (2006). *Acoustic Correlates of Word Stress in American English.* Doctoral dissertation, MIT, Cambridge.

Oller, D. K. (1973). The effect of position in utterance on speech segment duration in English. *Journal of Acoustical Society of America, 54,* 1235-1247.

O'Shaughnessy, D. (1979). Linguistic features in fundamental frequency patterns, *Journal of Phonetics, 7,* 119-145.

Pandey, P. C., Bhandarkar, S. M., Bachher, G. K., & Lehana, P. K. (2002). Enhancement of alaryngeal speech using spectral subtraction. *IEEE DSP 2002 (IEEE Aegean Island of Santorini), Vol. 2*, pp. 591-594.

Pratapwar, S. S., Pandey, P. C., & Lehana, P. K. (2003). Reduction of background noise in alaryngeal speech using spectral subtraction with quantile based noise estimation. *7th World Multiconference on Systemics, Cybernetics and Informatics (International Institute of Informatics and Systematics, Orlando)*, pp 408-413.

Qi, Y. Y., & Weinberg, B. (1991). Low-frequency energy deficit in electrolaryngeal speech. *Journal of Speech Hearing Research, 34*, 1250-1256.

Rothman, H. R. (1982). Acoustic analysis of artificial electronic larynx speech. In A. Sekey (Ed.), *Electroacoustic Analysis and Enhancement of Alaryngeal Speech* (pp.190-197), Springfield: Charles C Thomas Pub Ltd.

Saikachi, Y., Hillman, R. E., & Stevens, K. N. (2005). Analysis by synthesis of Electrolarynx speech. *Journal of Acoustical Society of America, 118.*

Sekey, A., & Hanson R. (1982). Laryngectomee speech support system with prosodic control. In A. Sekey (Ed.), *Electroacoustic Analysis and Enhancement of Alaryngeal Speech* (pp.166-183), Springfield: Charles C Thomas Pub Ltd.

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25(2)*, 193-247.

Slavin, C. S. & Ferrand, C. T. (1995). Factor analysis of proficient esophageal speech: toward a multidimensional model. Journal *of Speech and Hearing Research, 38*, 1224-1231.

Sluijter, A. M. C. (1995). *Phonetic Correlates of Stress and Accent.* Doctoral dissertation, Holland Institute of Generative Linguistics, Leiden.

Sluijter, A. M. C., Heuven, V. J. van, & Pacilly, J. J. A. (1997). Spectral balance as a cue in the perception of linguistic stress. *Journal of Acoustical Society of America, 101,* 503-513.

Takahashi, H., Nakao M., Kikuchi Y., & Kaga K. (2005). Alaryngeal speech aid using an intra-oral electrolarynx and a miniature fingertip switch. *Auris Nasus Larynx, 32,* 157-62.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychology Review, 34,* 273-286.

Torgerson, W. S. (1957). *Theory and Methods of Scaling.* New York, NY: John Wiley and Sons.

Uemi, N., Ifukube, T., Takahashi, M., & Matsushima, J. (1994). Design of a new electrolarynx having a pitch control function. *IEEE Workshop on Robot and Human,* 198-202.

van Geel, R. C. (1982). Semi-automatic pitch control for an electrolarynx. In A. Sekey (Ed.), *Electroacoustic Analysis and Enhancement of Alaryngeal Speech* (pp.190-197), Springfield: Charles C Thomas Pub Ltd.

Watson, J. W., & Schlauch, R. S. (2009). Fundamental frequency variation with an Electrolarynx improves speech understanding: a case study. *American Journal of Speech-Language Pathology, 18,* 162-167.

Weinberg, B., & Gandour, J. (1986). Prosody in alaryngeal speech. *Seminars in Speech and Language, 7,* 95-107.

Weiss, M. S., Yeni-Komshian, G. H., & Heinz, J. M. (1979). Acoustical and perceptual characteristics of speech produced with an electronic artificial larynx. *Journal of Acoustical Society of America, 65,* 1298-1308.

Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica, 49,* 25-47.

Winer, B. (1971). *Statistical principles in experimental design.* New York: MacGraw Hill.

Williams, C. E. & Stevens, K. N (1972). Emotions and Speech: Acoustical correlates. *Journal of Acoustical Society of America, 52,* 1238-1250.