Talker-to-Listener Distance Effects on Speech Production and Perception

Mark Clements, Kaustubh Kalgaonkar & Jonathan Kim
Center for Signal and Image Processing,
School of Electrical & Computer Engineering.
Georgia Institute of Technology, Atlanta, GA 30332

ABSTRACT

Simulating talker-to-listener distance (TLD) in virtual audio environments requires mimicking natural changes in vocal effort. Studies have identified several acoustic parameters manipulated by talkers when varying vocal effort. However, no systematic study has investigated vocal effort variations due to TLD, under natural conditions, and their perceptual consequences. This work examined the feasibility of varying the vocal effort cues for TLD in synthesized speech and real speech by (a) recording and analyzing single word tokens spoken at a range between 1 and 32 meters, (b) creating synthetic and modified speech tokens that vary in one or more acoustic parameters associated with vocal effort, and (c) conducting perceptual tests on the reference, synthetic, and modified tokens to identify salient cues for TLD perception. Measured changes in fundamental frequency, intensity, and formant frequencies of the reference tokens across TLD were similar to other reports in the literature. Perceptual experiments that asked listeners to estimate TLD showed that TLD estimation is most accurate with real speech; however significant standard deviations in the responses suggest that reliable judgments can only be made for gross changes in TLD.

I. INTRODUCTION

Current virtual audio environments can effectively simulate the direction to a sound source relative to the listener, but have difficulty simulating distance. Using distance cues such as overall level and direct-to-reverberant ratio requires the listener to have prior knowledge of the intensity of the source and/or the reverberance of the acoustic environment. One class of sounds with which listeners are very familiar is speech from a conversant. Listeners know from experience that a talker adjusts his or her vocal effort to compensate for the acoustic loss due to the distance separating them. A listener can use the learned acoustic correlates of vocal effort, together with the received sound level, to estimate the talker-to-listener distance (TLD). In order to exploit this prior knowledge of vocal effort cues for improving the simulation of talker distance in virtual audio environments, it would be necessary to have a means of imposing those cues onto either synthesized or processed speech.

Several recent investigations have sought to characterize the role of particular acoustic parameters in the perception of vocal effort as it relates to distance. These data were used as a basis for formulating the hypotheses described below. One set of experiments performed by Brungart & Scott[1] tested the role of talker speech intensity (production level) versus intensity at the listener (playback level) in a distance identification listening task. Talkers were recorded in an anechoic room with a microphone 1m from their mouth. They were instructed to produce single words and short phrases at selected SPL thresholds in 6 dB intervals. Then the speech was played back over headphones to a listener in a large, open field, where the listener had several distance markers serving as visual cues. They found that perceived TLD doubled with every 8dB increase in production level for levels greater than or equal to 66 dB SPL at 1m. Brungart et al.[2] then investigated how fundamental frequency (f0) affects the perception of distance by pitchshifting speech such that the “up-shifted” or “down-shifted” speech had a mean f0 that matched unmodified speech produced at a 6 dB higher or lower level, respectively. The listeners again performed a distance identification task with up-shifted, unmodified, and down-shifted speech. Their results indicated that down-shifted speech provides at least half of the perceived TLD change when compared to unmodified speech above 78 dB SPL at 1m, while up-shifted speech provides less than half of the perceived TLD change when compared to the unmodified speech.

Liénard & Di Benedetto[3] recorded isolated French vowels at three different indoor TLD’s: 0.4, 1.5 and 6 m. Vocal effort at each distance was first established by a qualitative mutual comprehension exercise, in which the listener (experimenter) and the talker engaged in a brief dialog. Their analysis concentrated on the spectral properties of the vowels that changed with vocal effort as a function of TLD. The data suggested that increases in fundamental frequency (f0), first formant frequency (F1), and first through third formant amplitudes (A1, A2, and A3) occur as the TLD increases.

Eriksson & Traunmüller[4] also used isolated vowels as stimuli, with the changes in vocal effort being varied by the distance between talker and experimenter, and evaluation by the experimenters’ judgments. Similar to Brungart’s work, they conducted distance identification listening tasks, but with two separate tasks for the listeners: (a) estimate the talker-to-receiver (communication) distance, where the listener in the perceptual experiment is not the receiver; and (b) estimate the talker-to-listener (listening) distance, where the recorded talker is speaking to the listener in the perceptual experiment. Their results suggested that listener’s memory is more robust in estimating the communication distance, which they believe is related to the talker’s vocal effort, than for estimating the listening distance, which they believe is related to the overall acoustic losses due to distance. They also noted increases in f0, F1 and decreases in spectral tilt (the relation of the formant amplitudes), which corroborated the results of Liénard & Di Benedetto.[3]

Traunmüller & Eriksson[5] investigated segmental and suprasegmental changes that occur with increasing vocal effort as their subjects spoke one sentence (in Swedish) over increasing TLD's in the set {0.3, 1.5, 7.5, 37.5, 187.5 m}. They reported results from 12 adults (six female, six male) and eight children (4 female, 4 male) showing increases in intensity, fundamental frequency (F1), and vowel durations as TLD increases. They sampled the largest range of TLD's of the studies reviewed, but with low resolution.

Tassa & Liénard[6] used speech modification techniques to transform speech produced with one level of vocal effort to another. Demonstrations, but no data, are available at their website suggesting that prosodic factors (e.g., intensity, pitch and segment durations) are much more important for transforming vocal effort than are spectral factors (e.g., formant frequencies or spectral tilt).

Other studies that manipulated vocal effort, but did not directly study the effect of TLD on vocal effort, used masking noise to elicit changes in subjects' vocal effort,[7,8] or asked the subjects to achieve quantitative[9] or qualitative[10] target levels (e.g., using a sound level meter or requesting "maximum vocal effort").

This work presents a systematic, comprehensive study of the effect of TLD on vocal effort by combining aspects of different past studies with logical extensions of their methods. Specifically, this work included

1) conducting the collection of the speech corpus in a large, open field, with actual talkers and listeners, to better capture the natural changes that occur in vocal effort with variations in TLD, unlike previous studies;

2) sampling the TLD space with more resolution than previous studies (e.g., Liénard & Di Benedetto[3]), while also sampling a large range of TLD like Brungart & Scott[1];

3) establishing comprehension at each TLD through a reading task requiring a qualitative response from an actual listener, similar to Liénard & Di Benedetto[3];

4) at each TLD, having the talkers speak a set of isolated words that contain a wide range of vowels and consonants;

5) maintaining the natural changes in vocal effort due to TLD through a two-alternative, forced-choice comprehension task for each word spoken; and

6) asking listeners to estimate the communication distance and the listening distance in separate perceptual tasks, similarly to Eriksson & Traunmüller[4].

This study used a formant-based speech synthesizer, and describes modifications to a standard vocoder algorithm used to produce synthetic speech and modified real speech with varied vocal effort intended to mimic changes in TLD. Lastly, results from preliminary perceptual studies of the real, modified, and synthesized speech are given

The project has three important tasks (a) reference recording (b) analysis of the data and (c) development of algorithms for modification of the vocal effort. The reference recordings were produced by the team at Sensimetrics. The pitch and formant tracking algorithms required for the analysis phase were developed at Georgia Institute of Technology. Two modification algorithms were generated (1) MELP vocoder based and (2) Sinusoidal Model based. Details about all the tasks will be provided in following sections.

II. METHODS

A. Reference Recordings

The reference recordings were conducted in a large, open farm field. Four participants, males (P1, P2) and females (P3, P4) each performed in talker-listener pairs: P1-P2, P2-P3, P3-P4, P4-P1. Speech tokens included 12 pairs of single words, taken from the Diagnostic Rhyme Test (DRT), repeated 3 times by the four talkers at 11 discrete talker-to-listener distances (TLD's) in the set {1, 1.4, 2, 2.8, 4, 5.6, 8, 11.2, 16, 22.4, 32 m}, for a total of 3,168 words. Using talker-listener pairs promoted natural changes in vocal effort with TLD, through both subjective feedback to the talker (e.g., "You're talking too quietly.") and objective feedback via a forced-choice word discrimination task for the listener. If a listener marked the incorrect word in a pair, the experimenter asked the talker to repeat that word once at the end of the word list for that distance. The corpus of DRT words (see Table I) was selected to maximize the sampling of consonant and vowel categories while minimizing the total number of words recorded.

Table I. The 12 word-pair subset of the DRT used as reference speech tokens.

beat – meat	daunt – taunt	rue – you
boast – ghost	dill – gill	sag – shag
boss – moss	fan – pan	sank – thank
caught – taught	mall – shawl	wall - yawl

A Marantz PMD-600 digital recorder captured the signals (Fs = 44.1kHz) from two Sennheiser MKE-2 microphones during the reference recordings. One microphone was worn on the talker's head and held at a fixed 15cm from the talker's mouth, while the other microphone was fixed on a stand 0.5m away from the listener, approximately at the listener's ear height, along the line between the talker and listener. The word order in the 24-word lists was randomized, both within and across TLD. Furthermore, to minimize vocal fatigue, the TLD performance order was randomized, and each talker-listener pair performed at only one TLD before switching to another pair.

Acoustic analysis of the head-mounted microphone signal involved locating the 2048-point (46.4 ms) window that contained the maximum energy during the vocalic portion of each DRT word. Over that window, the following parameters were calculated – overall intensity, average fundamental frequency (f0), first formant frequency (F1), second formant frequency (F2), third formant frequency (F3), first formant amplitude (A1), second formant amplitude (A2), third formant amplitude (A3), first harmonic amplitude (H1), and second harmonic amplitude (H2). The quantity H1-H2 was taken as a correlate of open quotient (OQ): the percentage of time that the vocal folds are open during one fundamental period, after Hanson & Chuang.[11] A linear regression was applied to the changes in each parameter across TLD, averaged over all 24 DRT words and all four talkers, except for f0 and F1 for which individual talker regressions were calculated.

Four hypotheses were tested through the analysis of the reference recordings. These hypotheses were based primarily on Brungart & Scott's finding of an 8dB increase in intensity per distance doubling for speech above 66 dB SPL at 1m, and Liénard & Di Benedetto's reporting of the average rate of increase with level for (a) f0 (5.1 Hz/dB), (b) F1 (3.5 Hz/dB), (c) A1 (1.1 dB/dB), and (d) A3 (1.3 dB/dB). The hypotheses are as follows:

1) First format amplitude A1 increases with vocal effort as a function of distance, with approximately an 8dB increase per distance doubling.

2) Fundamental frequency increases with vocal effort as a function of distance, with approximately a 40 Hz increase per distance doubling.

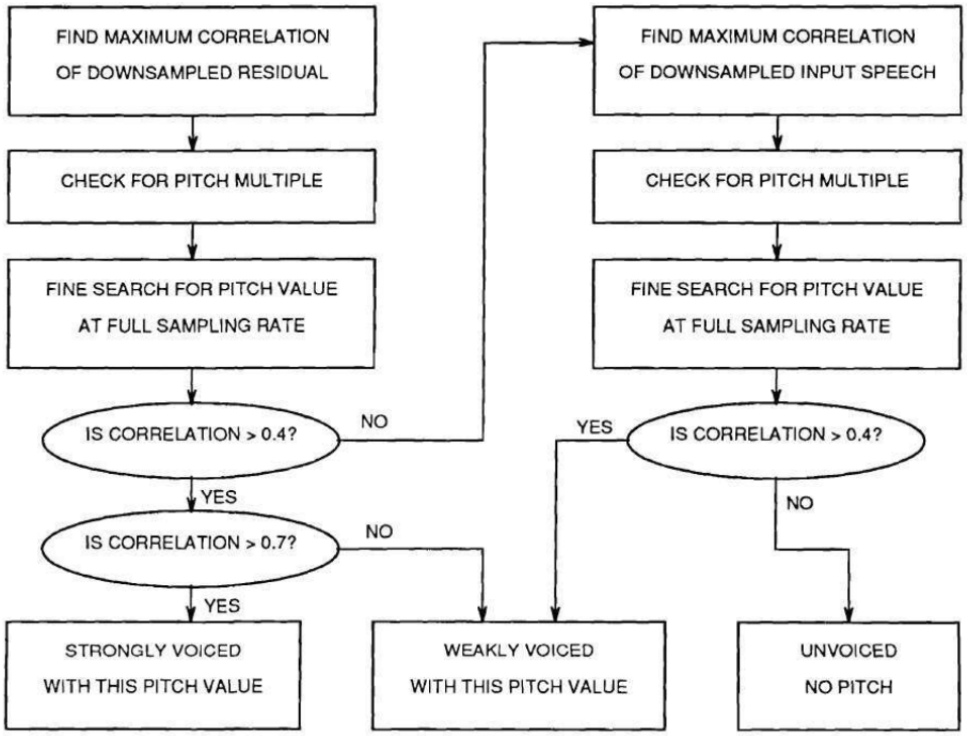
3) First formant frequency F1 increases with vocal effort as a function of distance, with approximately a 25 Hz increase per distance doubling.

4) Spectral tilt, measured as 3^{rd} formant amplitude minus 1^{st} formant amplitude (A3-A1) increases with vocal effort as a function of distance, with approximately a 2dB increase per distance doubling.

B. Speech Analysis Framework

1. Pitch Analysis

The pitch analysis algorithm was just a modified version of the standard MELP based pitch analysis algorithm. Pitch tracking can be briefly represented using the following flow chart.

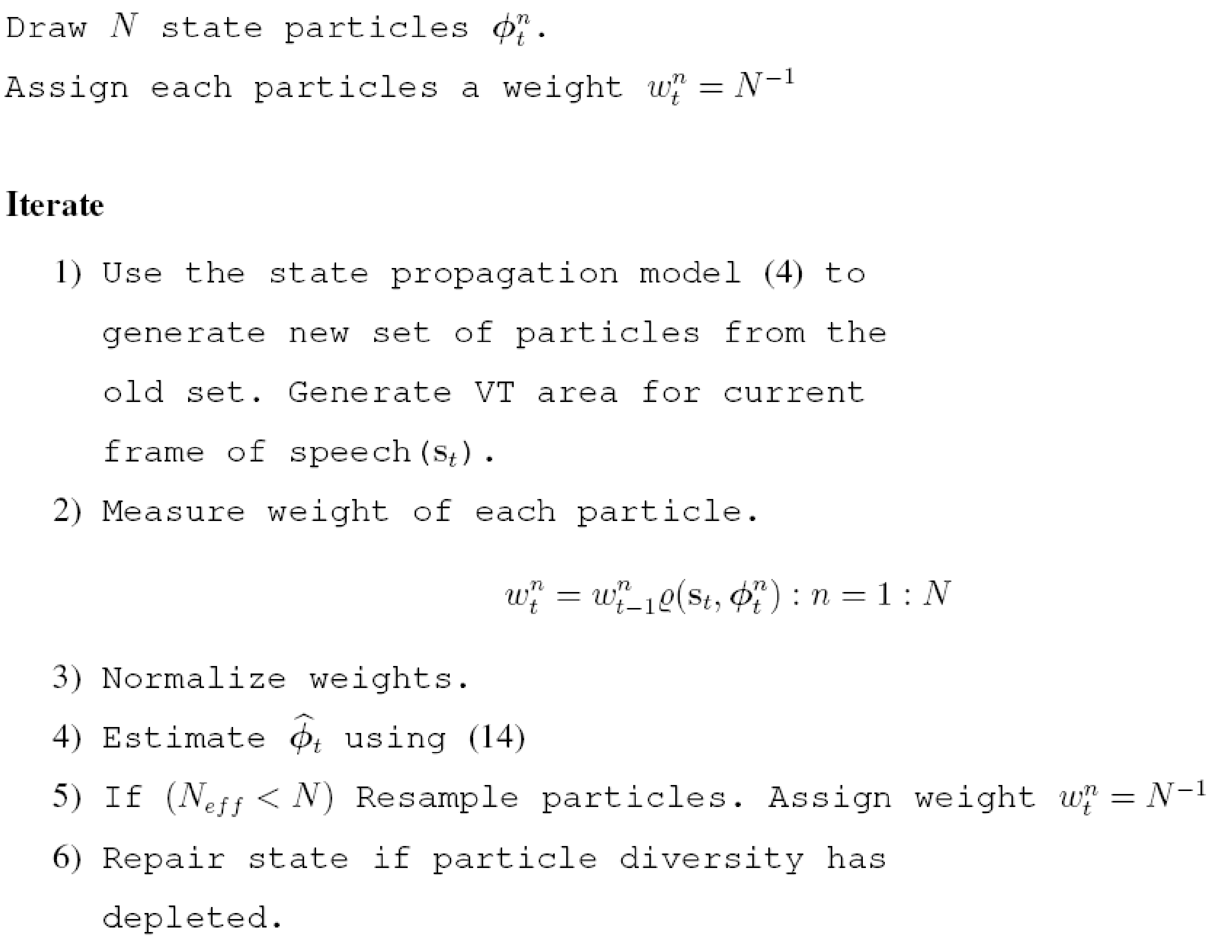

FIND MAXIMUM CORRELATION OF DOWNSAMPLED RESIDUAL
CHECK FOR PITCH MULTIPLE
FINE SEARCH FOR PITCH VALUE AT FULL SAMPLING RATE
IS CORRELATION > 0.4?
NO
YES
IS CORRELATION > 0.7?
NO
YES
STRONGLY VOICED WITH THIS PITCH VALUE
FIND MAXIMUM CORRELATION OF DOWNSAMPLED INPUT SPEECH
CHECK FOR PITCH MULTIPLE
FINE SEARCH FOR PITCH VALUE AT FULL SAMPLING RATE
YES
IS CORRELATION > 0.4?
NO
WEAKLY VOICED WITH THIS PITCH VALUE
UNVOICED NO PITCH

FLOWCHART 1 : Pitch Estimation Algorithm Flowchart.

2. Formant Tracking.

Two methods for formant tracking were developed. (1) using particle filters and (2) using nonlinear optimization.

a. **Formant Tracking using Particle Filters:** A detailed explanation of the formant tracking using particle filter can be found in [1] Kalgaonkar and Clements. The algorithm in brief is stated below:



```
Draw $N$ state particles $\phi_t^n$.
Assign each particles a weight $w_t^n = N^{-1}$

Iterate
    1) Use the state propagation model (4) to
       generate new set of particles from the
       old set. Generate VT area for current
       frame of speech($\mathrm{s}_t$).
    2) Measure weight of each particle.
            $w_t^n = w_{t-1}^n \varrho(\mathrm{s}_t, \phi_t^n) : n = 1 : N$
    3) Normalize weights.
    4) Estimate $\widehat{\phi}_t$ using (14)
    5) If $(N_{eff} < N)$ Resample particles. Assign weight $w_t^n = N^{-1}$
    6) Repair state if particle diversity has
       depleted.
```

Where $\widehat{\phi}_t = \int \phi_t \sum_{n=1}^{N} w_t^n \delta(\phi_t - \phi_t^n) d\phi_t = \sum_{n=1}^{N} w_t^n \phi_t^n$ is the estimate of the formant bandwidth and formant frequency.

b. Formant Tracking using Non-Linear Optimization : The formant estimate is obtained by minimizing the squared error between the vocal tract area obtained from the formant estimate and the actual Vocal tract area obtained using the method described in [2]. The optimization technique used is a modified version of Newton’s method, where the inverse of the Hessian is obtained using BFGS update algorithm. The details can be found in the attached code. The optimization based method provides better estimates than the particle filter based formant estimator and is used for analysis.

C. Speech Modification and Synthesis Framework

1. High-Level synthesis (HLsyn)

HLsyn a quasi-articulatory speech synthesis system, developed at Sensimetrics Corporation.[12,13] is driven by a 13-component vector updated every 5ms. The HLsyn concept is called "quasi-articulatory" because it is based on the observation that many parameters needed to control a formant synthesizer are not independent, but are constrained by the anatomy and physiology of the human speech production system. HLsyn incorporates many of these constraints in a higher-level control system layered on top of a conventional formant synthesizer (KLsyn) that generates the output speech. Thirteen parameters control HLsyn, and are transformed into the more than 40 KLsyn[14] parameters through a set of mapping relations.

The 13 components are the fundamental frequency, the first four formants, and eight more parameters related to the physiology of speech production (e.g., the area of the velopharyngeal port *(an)*, the area of the opening at the lips *(al)*, the area of the tongue blade constriction *(ab)*, etc.). The subglottal pressure parameter *(ps)* can effect changes in f0 and overall intensity, both of which have already been identified as correlates of vocal effort.[1,2,3,4,14,15]

2. Synthetic Speech

Synthetic speech tokens were generated using the HLsyn speech synthesizer (Sensimetrics Corp.), a quasi-articulatory formant-based synthesizer. The acoustic parameters f0, overall intensity, A1 and spectral tilt (TL) were manipulated in the synthetic speech tokens to investigate their effect on the perceived TLD. The HLsyn stimuli modeled only one talker, and were not intended to sound similar to any of the participants. One set of synthetic speech tokens only varied in f0, with the intent of identifying the relative importance of changing f0 to the other parameter changes. Similar individual changes in A1 were not tested perceptually because its changes resulted in speech that sounded unnatural, so the changes in TL and A1 were combined into a second synthetic speech token set.

3. MELP Vocoding

The MELP (Mixed Excitation Linear Prediction) coder is a purely parametric method that lends itself to high quality analysis/synthesis in the absence of quantization. It is the basis for the US DoD standard FS-1017 (MIL-STD-3005)[16] and the recently adopted NATO standard (NATO STANAG 4591). Its novelty is based on three innovations: (a) a highly accurate pitch extraction algorithm that has subsample precision, (b) a periodic excitation waveform that has both pulse shaping and a natural jitter, and (c) an overall excitation pattern that allows mixing periodic and non-periodic characteristics in a frequency-dependent fashion. Its parametric design is sufficiently modular so as to allow changing the excitation, the vocal tract filtering, or both. The speech modifications is carried out in two stages, first the incoming speech is parameterized, modifications to these parameters (LPC polynomial, excitation) are made to change the formant frequencies and amplitudes, spectral tilt and pitch. In the second stage speech is synthesized using these new parameters.

4. Modified Speech Using a MELP Vocoder

Modified speech tokens were produced using a modified version of the MELP vocoder described above - altered to allow manipulation of f0, overall intensity, A1, and spectral tilt. Changing the f0 of the MELP-vocoded speech is straightforward, as f0 is a parameter of the vocoder that can be directly modified. However, due to inherent limitations in the standard

MELP vocoding algorithm, increasing f0 above 200 Hz was not possible. This limitation is discussed more in Section III.B. The reference recordings made at TLD = 1m were processed by the vocoder to simulate TLD's of 4m, 8m, 16m, and 32m.

Although the parameters of the MELP vocoder allowed independent control of the vocal source and filter, the formant amplitudes could not be directly modified from the LPC polynomial *A(f)*. To overcome this limitation, the formant bandwidths were modified to indirectly change their amplitudes. To modify the formant bandwidths, each formant was modeled as a two pole digital resonator with a pair of complex conjugate poles located near the unit circle:

$$p_{1,2} = re^{\pm j\omega_0} \tag{1}$$

The amplitude of the transfer function of such a digital resonator is given by Eq. (2), after Proakis & Manolakis:[17]

$$\left|H(\omega_0)\right| = \frac{1}{(1-r)\sqrt{1+r^2-2r\cos(2\omega_0)}} \tag{2}$$

To manipulate the amplitude $\left|H(\omega_0)\right|$ independently of the pole frequency, the pole radius *r* can be changed. For *r* ≈ 1, changing *r* will also affect the bandwidth of the resonator B_{ω_0} according to Eq. (3):

$$B_{\omega_0} \approx 2(1-r) \tag{3}$$

Oppenheim & Schafer[18] have shown that in the LP spectrum, a formant bandwidth B_k can be approximated with Eq. (4) below, which relates the group delay to the bandwidths for single pole systems:

$$B_k \approx \frac{-1}{\pi}\left[\ln\left(\frac{\partial\angle A(f)}{\partial f}\right) - \ln\left(2\pi + \frac{\partial\angle A(f)}{\partial f}\right)\right]\Bigg|_{F_k} \tag{4}$$

where *A(f)* is the LP spectrum, given by

$$A(f) = 1 - \sum_{k=1}^{P} a_k e^{-j2\pi fk} \tag{5}$$

the derivative of *A(f)* is given by

$$\frac{\partial A(f)}{\partial f} = j2\pi \sum_{k=1}^{P} a_k e^{-j2\pi fk} \tag{6}$$

and the derivative of $\angle A(f)$ is given by

$$\frac{\partial \angle A(f)}{\partial f} = \operatorname{Im}\left\{ \frac{\frac{\partial A(f)}{\partial f}}{A(f)} \right\} \tag{7}$$

Setting Eq. (3) equal to Eq. (4) provides a method for approximating the formant pole radius *r* without explicit root solving as expressed in Eq. (8):

$$r \approx 1 + \frac{1}{2\pi}\left[\ln\left(\frac{\partial \angle A(f)}{\partial f} \right) - \ln\left(2\pi + \frac{\partial \angle A(f)}{\partial f} \right) \right]\Bigg|_{F_k} \tag{8}$$

Thus, the algorithm to alter the amplitude of a formant in an LPC spectrum can be stated as

1) Approximate the radius of the formant *r* using Eq. (8) above.
2) Choose a new radius $\tilde{r} = r + \Delta r$ to model the desired formant amplitude $\left|\tilde{H}(\omega_0)\right|$ using Eq. (2).
3) Calculate the new formant bandwidth using Eq. (3).
4) Use the LSP-based formant bandwidth modification method described by Morris & Clements[19] to obtain the new LPC coefficients.

Spectral tilt and gain were changed by manipulating several formant amplitudes simultaneously. FIG. 1 shows two spectral results of using the above algorithm to modify formant amplitudes.

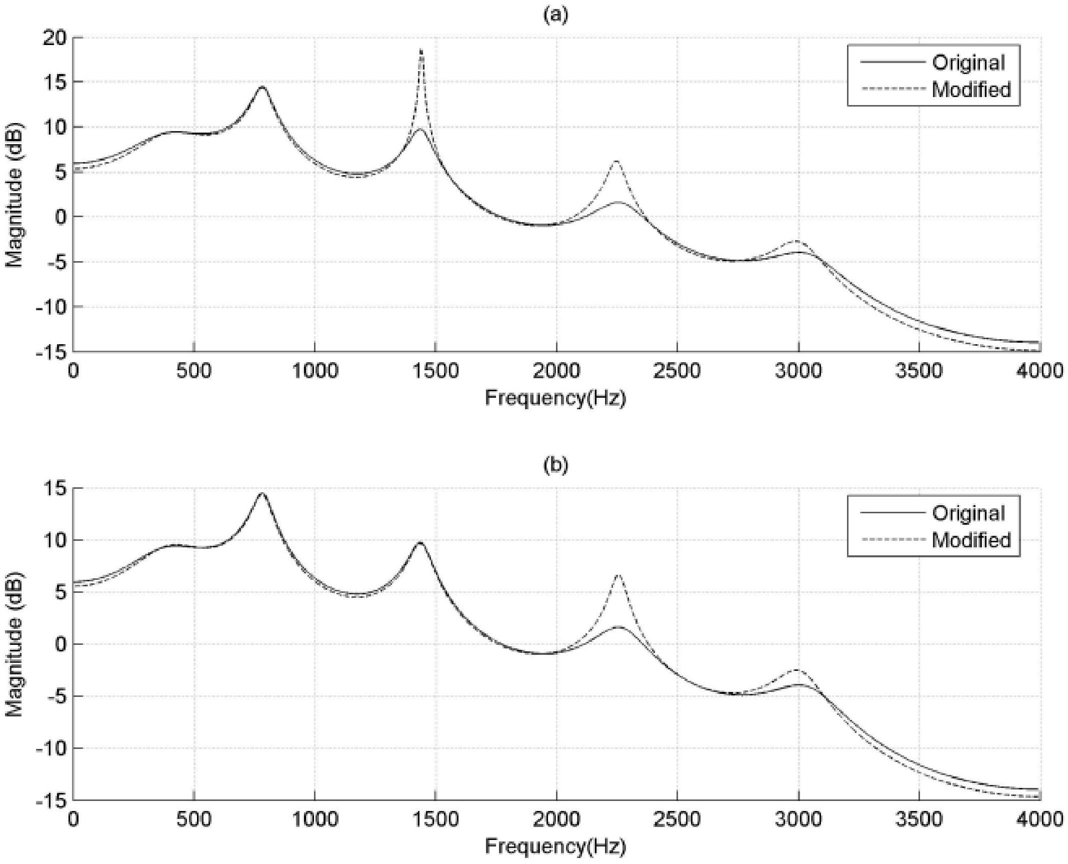
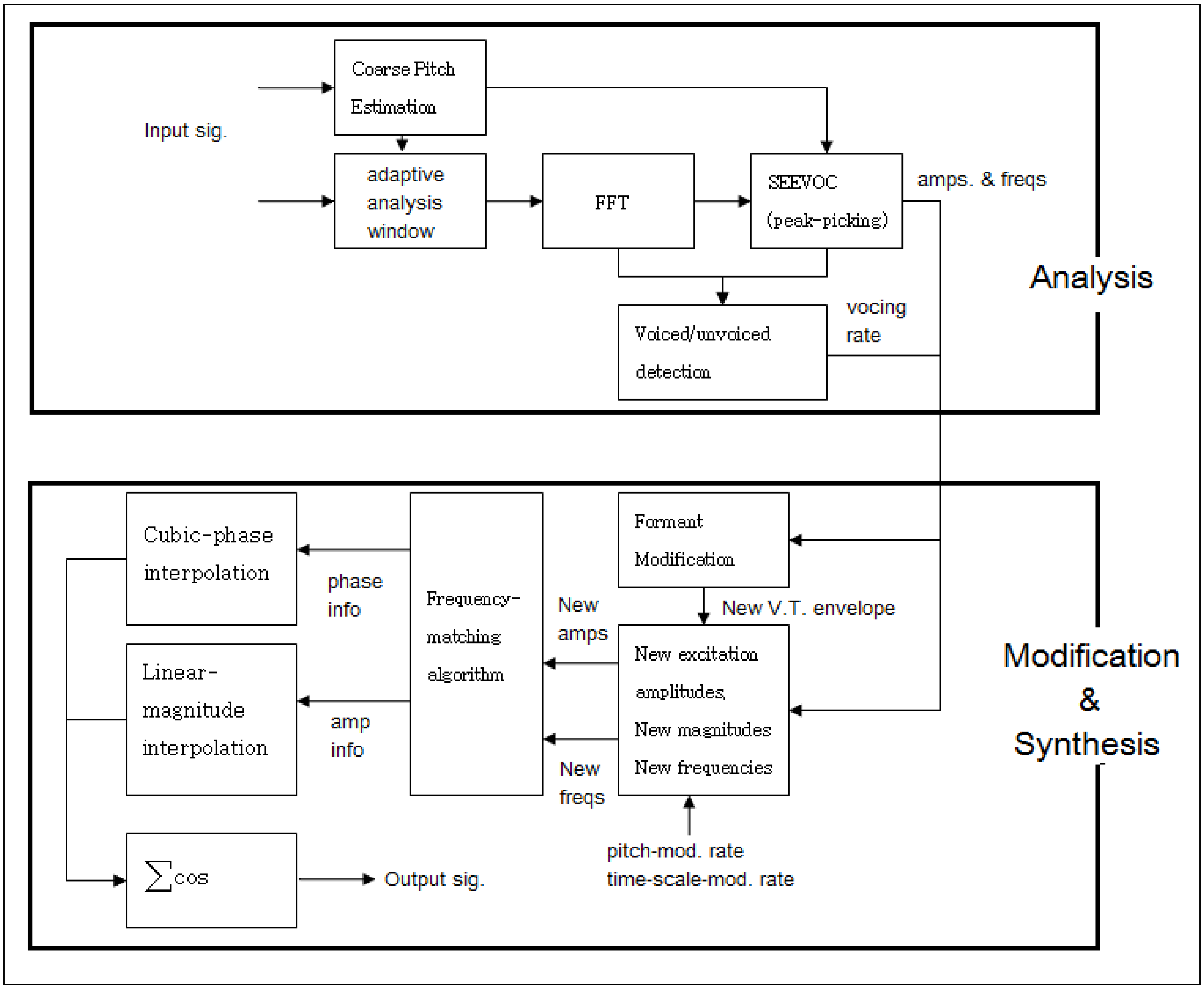


FIG. 1. Spectra for original (solid) and modified (dashed) LPC polynomials, showing increases in (a) A2, A3, and A4 by 8, 4 and 2 dB respectively; and (b) A3 and A4 by 4 and 2 dB.

5. Overview of STC

STC can be divided into analysis and synthesis parts. The STC algorithm developed by Georgia Tech is based on Quatieri's method. A block diagram of the baseline STC is illustrated in the figure below.



Block diagram of the baseline sinusoidal analysis/synthesis system.

In order to modify our input speech signal based on the TLD, our STC system performs pitch, time-scale, and formant modifications. An important property of the system is its capability of performing all these three modifications with time-varying rates of change.

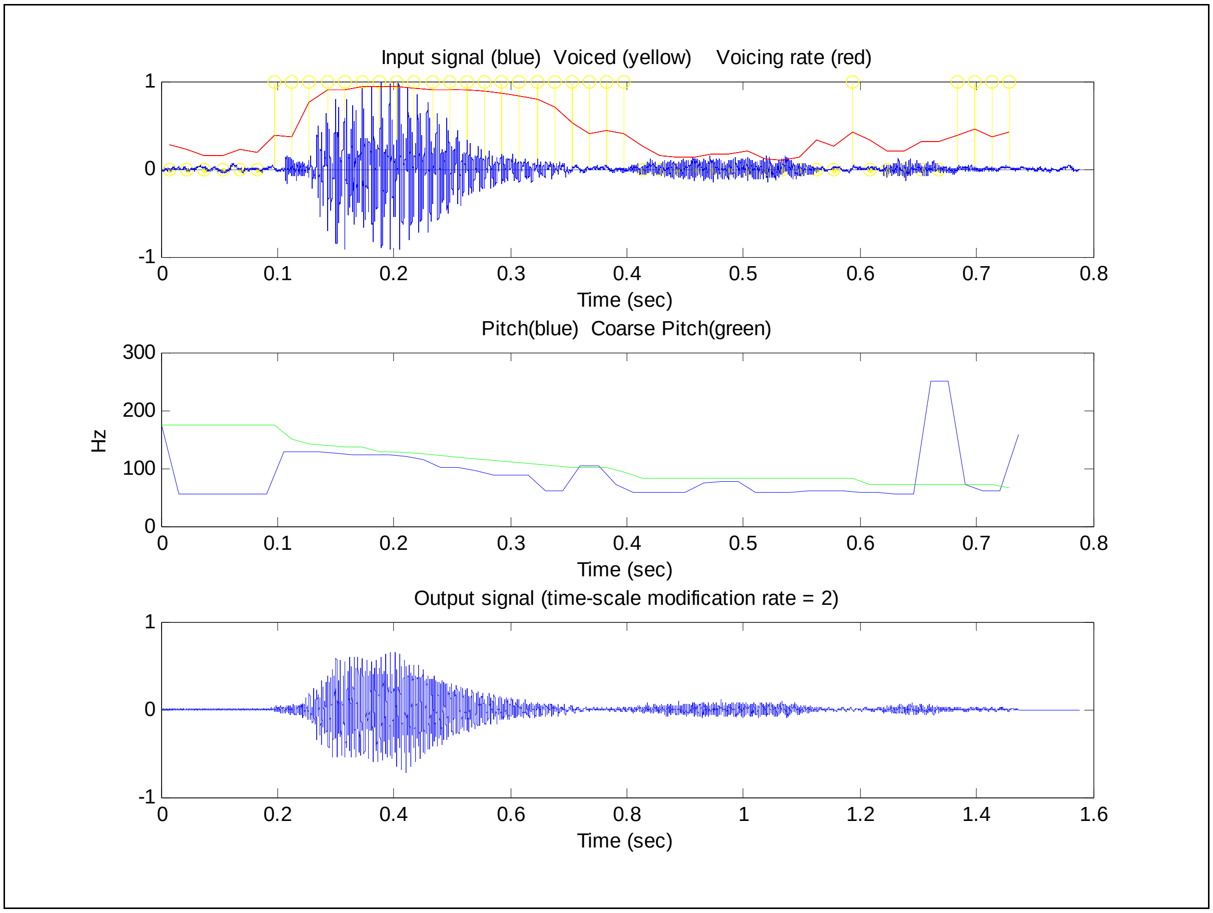
As shown in the block diagram, the system can be divided into three parts: analysis, modification, and synthesis. For our scenario, the analysis part is at the talker's end and the parameters are transmitted to the listener's end for the modification and the synthesis. The passing parameters from talker to listener are complex amplitudes, their corresponding frequencies, and the voicing rate. It is tested that ~100 amplitudes are enough for the sample frequency of 44.1 KHz.

At the listener's end, it requires the following parameters.

1) Amplitudes, frequencies, and voicing rate from the talker's end
2) Pitch-modification rate, time-scale-modification rate, and formant-amplification rate based on the TLD.
3) Since the system is using the adaptive analysis window based on the coarse pitch of the talker, the output speech is synthesized with different synthesis window length for each frame. The length of analysis window is also needed to be transmitted from the talker's end to the listener's end.

5.1 Time-Scale Modification

The goal of time-scale modification is to maintain the perceptual quality of the original speech while changing the apparent rate of articulation and the shape of the vocal tract. The case time-scale modification rate (ρ)>1 corresponds to slowing down the rate of articulation by means of a time-scale expansion. When ρ>1 the modified speech signal will speed up while keeping other speech properties.

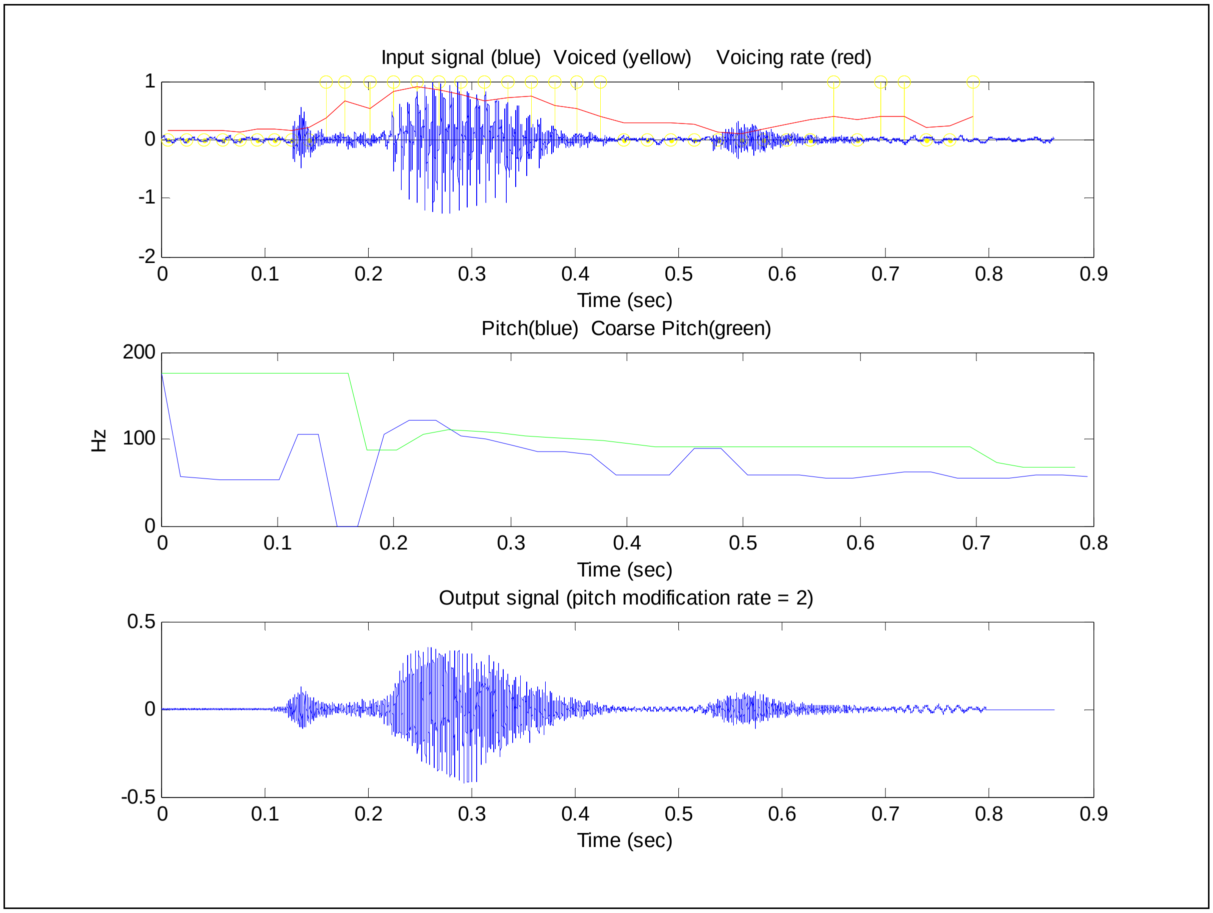


Example of time-scale modification with ‘p1_d1_ghost_1.wav’

The figure above shows an example of the time-scale modification with STC. The input speech signal is chosen from a male talker with TLD=1m. In the second plot, two different pitch estimations are plotted. The one with blue shows the instant pitch for each frame whose window length was about 30ms. One of the challenges of STC is that it requires the coarse pitch estimation. The coarse pitch estimation does not necessarily need to be precise, but somewhere between smooth and accurate. The inaccurate coarse pitch estimation due to the doubling and halving ambiguity produces notable artifacts in the synthesized signal. In order to remove this type of artifacts, the coarse pitch is estimated by averaging the previous voicing frames. This coarse pitch is plotted in green in the figure above.

5.2 Pitch Modification

The goal of the pitch modification is to modify the excitation function of the speaker while preserving the shape of the vocal tract and the duration of the signal. In our Phase I, it was shown that fundamental frequency (pitch) increases with vocal effort as a function of distance. Approximately 40% increase per distance doubling for speech above 66 dB SPL at 1m. When the pitch-modification rate β>1, the synthesized speech signal has a higher pitch than the original input signal. The sine-wave frequencies are scaled as $\omega_k' = \beta\omega_k$.

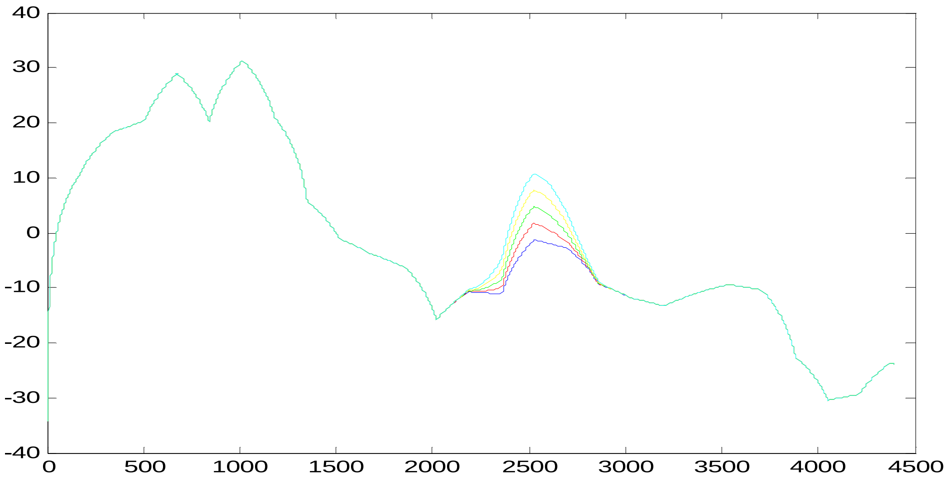


Example of pitch modification with 'p1_d1_taught_1.wav'

5.3 Vocal Tract Modification

It was shown that the first formant amplitude and the spectral tilt increase with vocal effort as a function of distance. Approximately 8 dB of the first formant amplitude and 3 dB of the spectral tilt increase per distance doubling for speech above 66 dB SPL at 1m. In order to amplify the formant amplitudes, four Gaussian curves are used. The each Gaussian curve has its mean at a formant frequency.

The input signal, “swing your arm as high you can.” is from TIMIT. The figures below are taken from a segment “arm” from the input signal. The formant frequencies from Deng’s database (MSR-UCLA VTR-FORMANT DATABASE) are used.

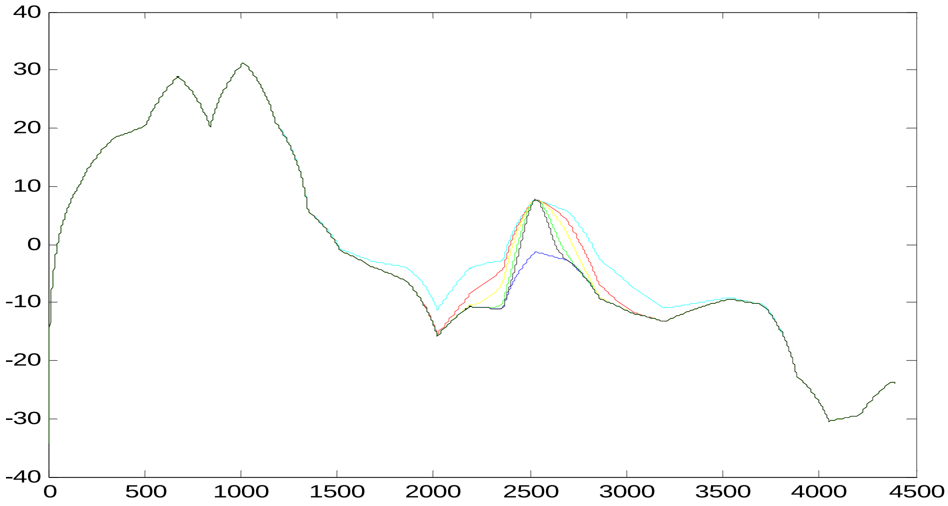


The figure above shows the third formant modifications (amplification) by 3dB(red), 6dB(green), 9dB(yellow), 12dB(cyan), and the original spectral envelope estimated by SEEVOC algorithm (blue).

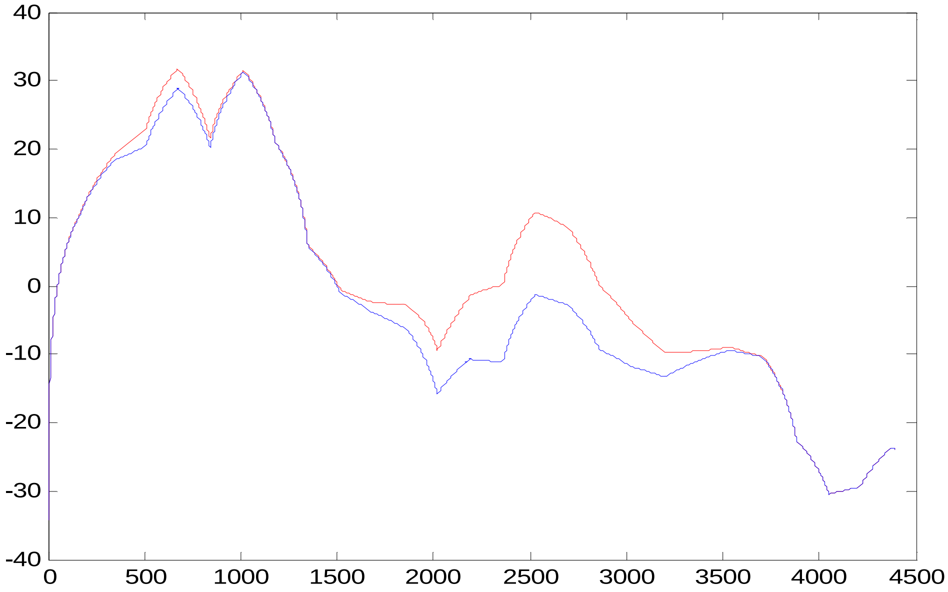
The formant frequency bandwidths are initially calculated by finding the minimum distances with its neighboring formant frequencies. For example if F1=400Hz, F2=1000Hz, F3=3000Hz, F4=4500Hz, then bandwidths are 400Hz(=min(400,600)), 600Hz(=min(600,2000)), 1500Hz(=min(2000,1500)), and 500Hz(=min(1500,500)). The maximum possible formant frequency is assumed to be 5000Hz.

With these formant frequencies and the bandwidths, the new spectral envelope is designed using Gaussian curve. For four formant frequencies, four Gaussian curves are designed and added. Each Gaussian curve has its mean at the given formant frequency and its sigma (variance of a Gaussian) in term of the bandwidth.

The variable in MATLAB code, fmbw is used for this sigma. The sigma of the Gaussian curve is estimated by dividing the bandwidth with fmbw, so that bigger fmbw gives the narrower curve.



The figure above shows the third formant modifications with varying the bandwidth. The original spectral envelope is plotted blue. The new spectral envelopes are designed for third formant modification with 9dB amplification and fmbw=3 (cyan), fmbw=6 (red), fmbw=9 (yellow), fmbw=15 (green), and fmbw=20 (black).



The figure above shows that how this formant modification method preserves other formants while modifying the desired formants. The figure above is amplifying the first formant with 3dB and the third formant with 12dB (fmbw=3) while second and the fourth formants stay unchanged.

III. RESULTS & DISCUSSION

A. Reference Recordings

Of the parameters measured for each word (see Section II.A.), the only parameters found to vary systematically with TLD were overall intensity (highly correlated with A1), f0, and spectral tilt (A3-A1).

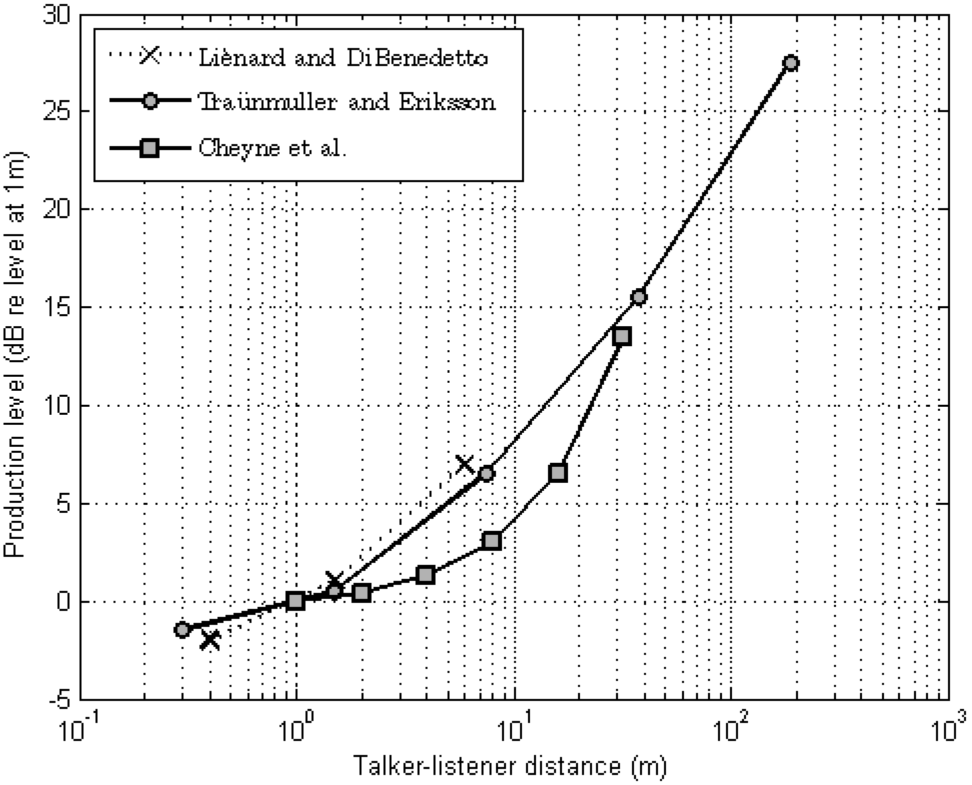


FIG. 2. Production level as a function of TLD (m): mean data from this study, Liénard & Di Benedetto,[3] and Traunmüller & Eriksson,[5] normalized to the level at TLD = 1m.

The change in overall intensity versus TLD is shown in FIG. 2, with a comparison of this study's data to that of Liénard & Di Benedetto[3] and Traunmüller & Eriksson.[5] The data was normalized to TLD = 1m for the comparison. Overall the data from the three studies display similar increases in intensity with TLD, a shallower increase of intensity than would be required to simply compensate for the 6dB per TLD doubling acoustic loss. Furthermore, the 8dB increase in production level per TLD doubling noted by Brungart & Scott[1] over their 0.25m to 64m range, and used as the basis for the second hypothesis above, was not observed in our data except between the 16m and 32m distances where the average change was 7dB. Differences in methodology may explain this discrepancy, as reference recordings used actual talkers and listeners with the goal of comprehension at each distance. In contrast, Brungart & Scott used talkers in an anechoic chamber with a production level goal, then played the recorded speech to listeners in a field with markers along the 0.25m to 64m range.

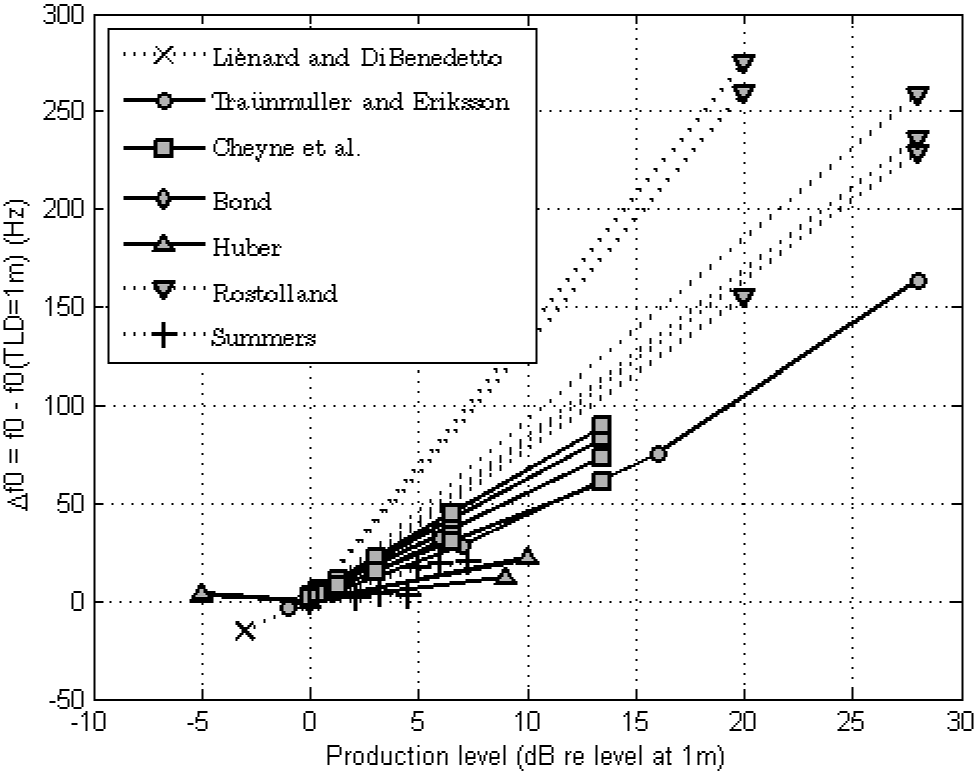


FIG. 3. Change in fundamental frequency (f0) relative to f0 at TLD = 1m as a function of normalized production level. This study's data compared to previous studies.

Fundamental frequency changes with production level are displayed in FIG. 3 for the individual talkers in this study, as well as individuals from previous studies[8,9,10] and mean data from previous studies.[3,5,7] The changes are averaged over the DRT words spoken. For comparison, the data has been normalized to the f0 and production level at TLD = 1m. The increases seen in f0 with TLD from this study were not as large as expected based on the results of Brungart et al.[2] and Liénard & Di Benedetto,[3] which led to the second hypothesis. Even at the largest distance doublings, the percent change in f0 is only about one-half of what was hypothesized. However, the P1 through P4 data fall within the range bounded by the previous studies shown, suggesting that basing the second hypothesis on extrapolating the f0 changes noted by Liénard & Di Benedetto[3] using the perceptual data of Brungart & Scott[1] overestimates the actual increases in f0 with TLD.

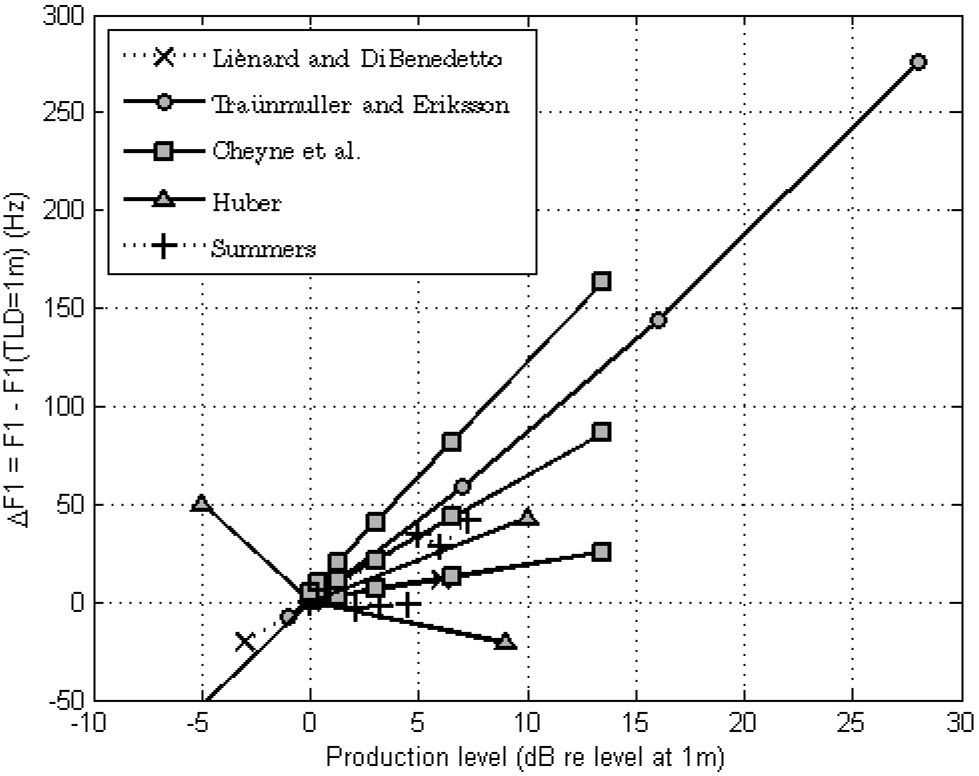


FIG. 4. Change in first formant frequency (F1) relative to F1 at TLD = 1m as a function of normalized production level. This study's data compared to previous studies.

First formant frequency (F1) changes with TLD were not consistent across talkers, similar to the f0 changes, as shown in FIG. 4. Again the data are normalized to the TLD = 1m condition, and the changes in F1 appear as a function of production level for ease of comparison with other investigators' data. Note that the data for P2 and P3 were so nearly identical that their square markers overlap in FIG. 4. The maximum change in F1 for a doubling of TLD ranged from 12 Hz (P1) to 82 Hz (P4), bracketing the hypothesized change of 25 Hz (based on the work of Eriksson & Traunmüller[4], Liènard & DiBenedetto[3], and Traunmüller & Eriksson[5]). However, those larger changes in F1 only occurred for the larger TLD values. Talkers P2 and P3 exceeded the hypothesized 25 Hz change for the 11.2m to 22.4m and 16m to 32m doublings, while talker P4 also exceeded that change for the 5.6m to 11.2m and 8m to 16m doublings.

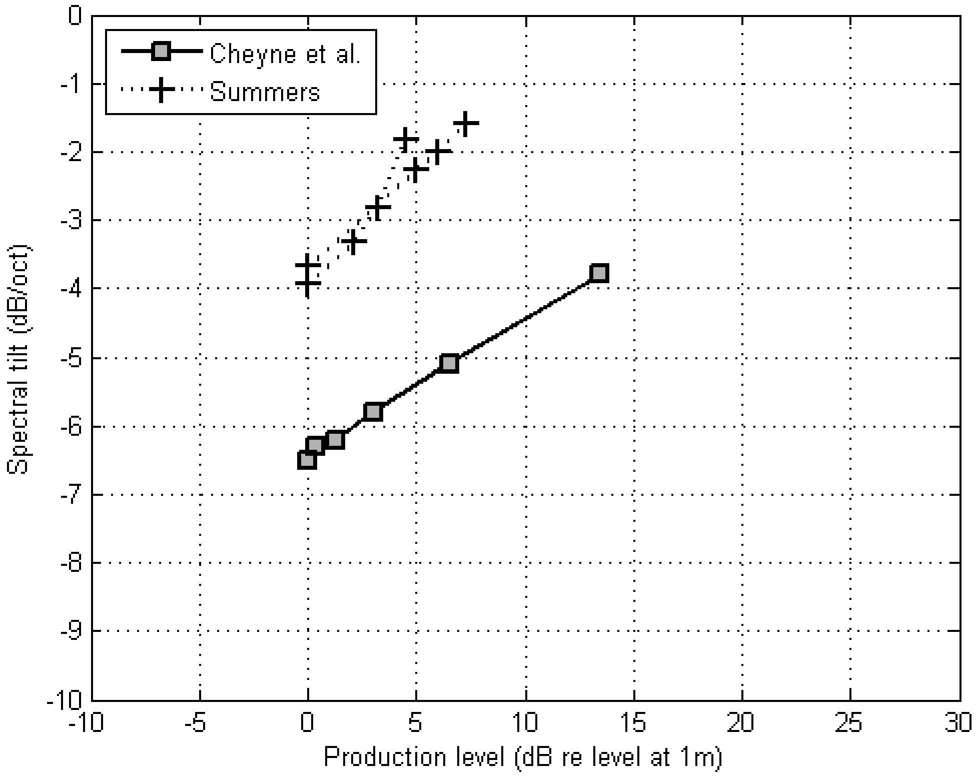


FIG. 5. Spectral tilt (TL) as a function of normalized production level. This study's data compared to data of Summers et al.

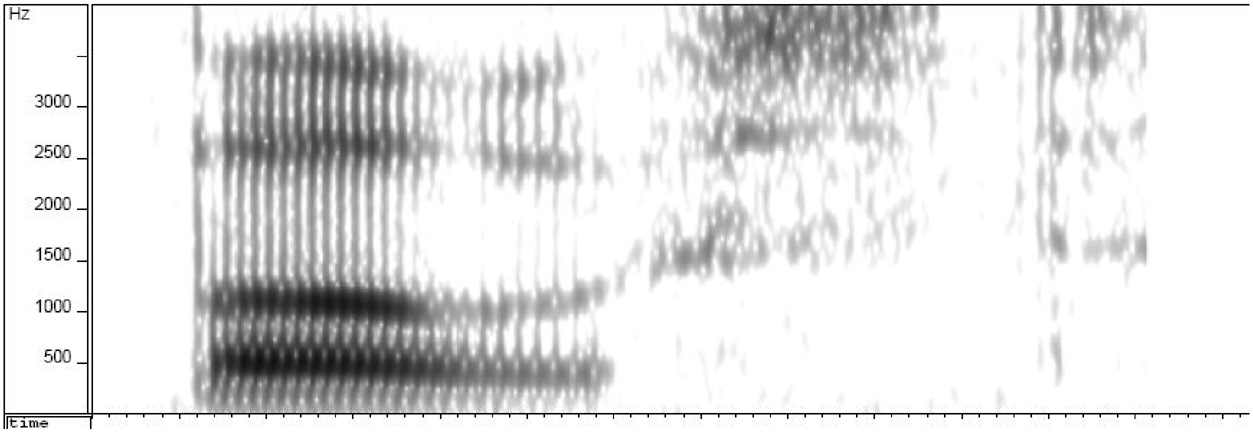
The spectral tilt (TL), defined as the third formant amplitude minus the first formant amplitude (TL = A3 - A1), did increase with TLD but not at the rate hypothesized above. FIG. 5

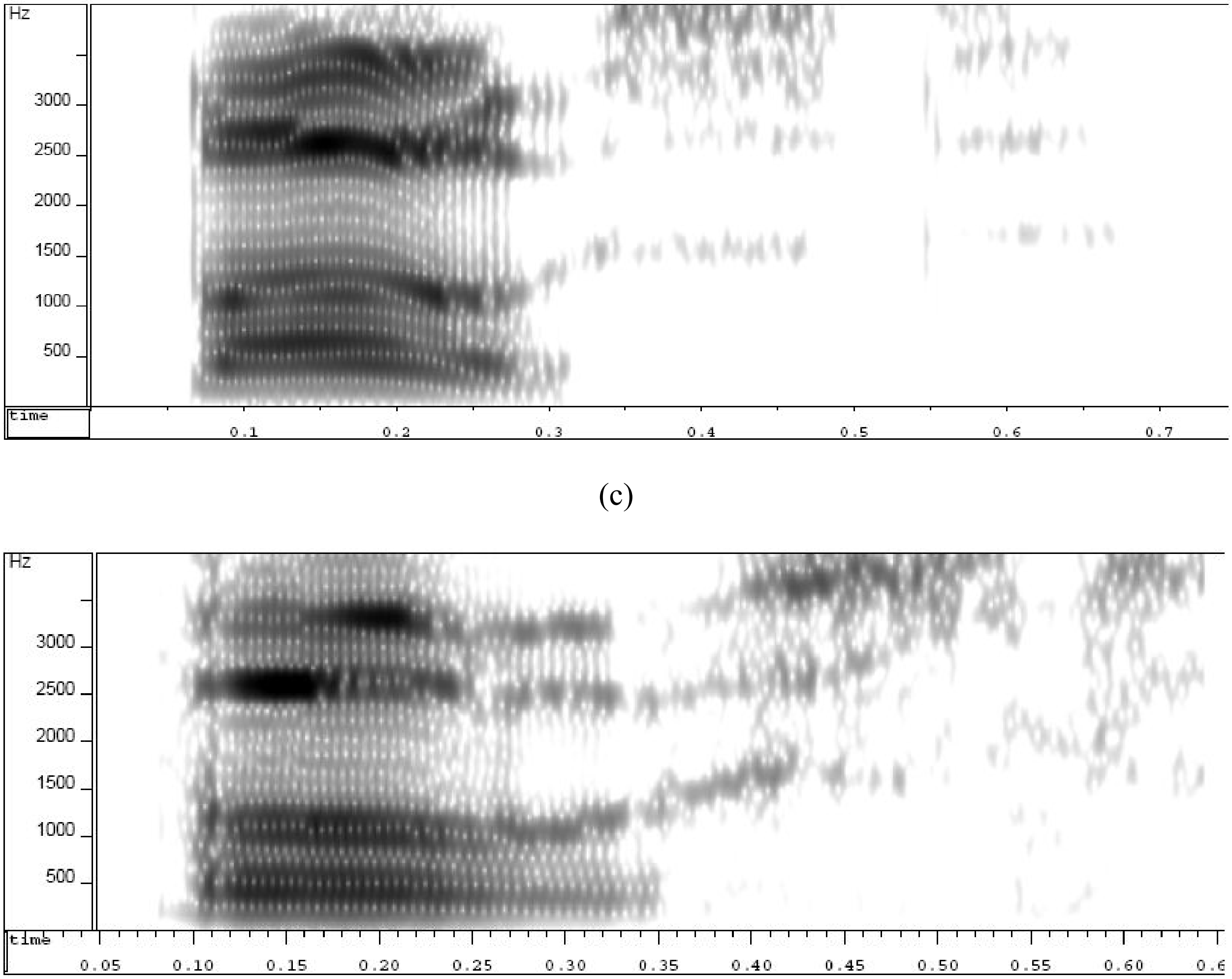
shows the mean value of spectral tilt versus the production level, again with the production level normalized to the level at TLD = 1m. The hypothesized 2dB increase per distance doubling, based on the formant amplitude changes noted by Liénard & DiBenedetto[3], was only nearly approached at the final TLD doubling from 16m to 32m. Compared to the data of Summers et al.,8 TL increased more gradually with production level, suggesting that the task difference – increasing vocal effort to overcome noise rather than distance – may lead the talkers to employ different strategies acoustically.

B. MELP Vocoded Speech Tokens

The modified MELP vocoder processed some of the reference recordings from TLD=1m to produce speech tokens that mimicked the changes in f0 and intensity that occurred for TLD's of 4, 8, 16, and 32m. Example spectrograms of original and modified speech tokens for the word "mall" appear in Figure 6.

(a)




(b)

(c)

FIG. 6. Spectrograms of the word "mall" from the reference recordings at TLD = 1m (a) and TLD = 32m (b), and from the modified MELP vocoder synthesizing TLD = 32 m from (a).

In MELP, fundamental periods are restricted to fall within a fairly narrow range. This range does not entirely include the f0 values measured from the reference recordings. Modifying the MELP framework to expand the f0 range is non-trivial, as the framing, short-term predictors, long-term predictors, harmonic voiced-unvoiced assignments, and pulse dispersion would all be affected. In this work, only target f0 values that were in the standard MELP range were included. For example, if the f0 of P4 saying "wall" at TLD = 32m exceeded the maximum MELP f0, then the sample of P4 saying “wall” at TLD = 1m was not used as an input to the MELP vocoder for creating modified tokens. Using that limitation, for male speakers, f0 was increased by 30 Hz to model TLD = 16m, and 65 Hz for TLD = 32m. For female speakers, f0 was increased by 20 and 40 Hz for TLD = 16 and 32 m, respectively.

Although MELP controls available to us could be used to modify formant bandwidths, experience with modifying the formants suggests that the bandwidths of all poles must be modified to effect the desired change in amplitude of even one formant. This likely follows from the anatomical changes used to alter one formant’s bandwidth or amplitude, which will affect the other formants (e.g., stiffening the vocal tract walls). Further, spectral tilt also affects the formant amplitudes. For the set of MELP tokens with modified spectral tilt and gain, the spectral tilt was increased by 2 dB/oct to synthesize vocal effort for TLD = 16m, and 4 dB/oct for TLD = 32m. The spectral gain was increased by 8 and 15 dB for synthesizing the vocal effort for TLD = 16m and TLD = 32m, respectively. Unfortunately some formant and f0 modifications, particularly for the higher f0 examples, resulted in modified speech that sounded unnatural, possibly confounding the perceptual results.

C. Perceptual Testing

The perceptual testing presented subjects with randomized speech tokens from one of the three sets of tokens: the reference recordings, synthetic speech, or modified speech. Subjects were asked to identify the distance between the talker and the conversant in two assumed contexts: first, where the listener is the conversant (the “two-party” case), a situation in which the speech signals contain the full set of distance-related acoustic cues; and second, where a third party is the conversant at a variable distance while the listener is at a fixed distance of 1m from the talker (the “three-party” case), a situation that conveys vocal effort cues but not the level variation due to distance. The two questions were asked separately, that is, after one subject had listened to an entire set of reference tokens and judged the 2-party distance, they listened to the set again without attenuation due to distance and were asked to estimate the 3-party distance. For audio samples of the speech tokens, visit the website www.sens.com/distance_cues/samples.

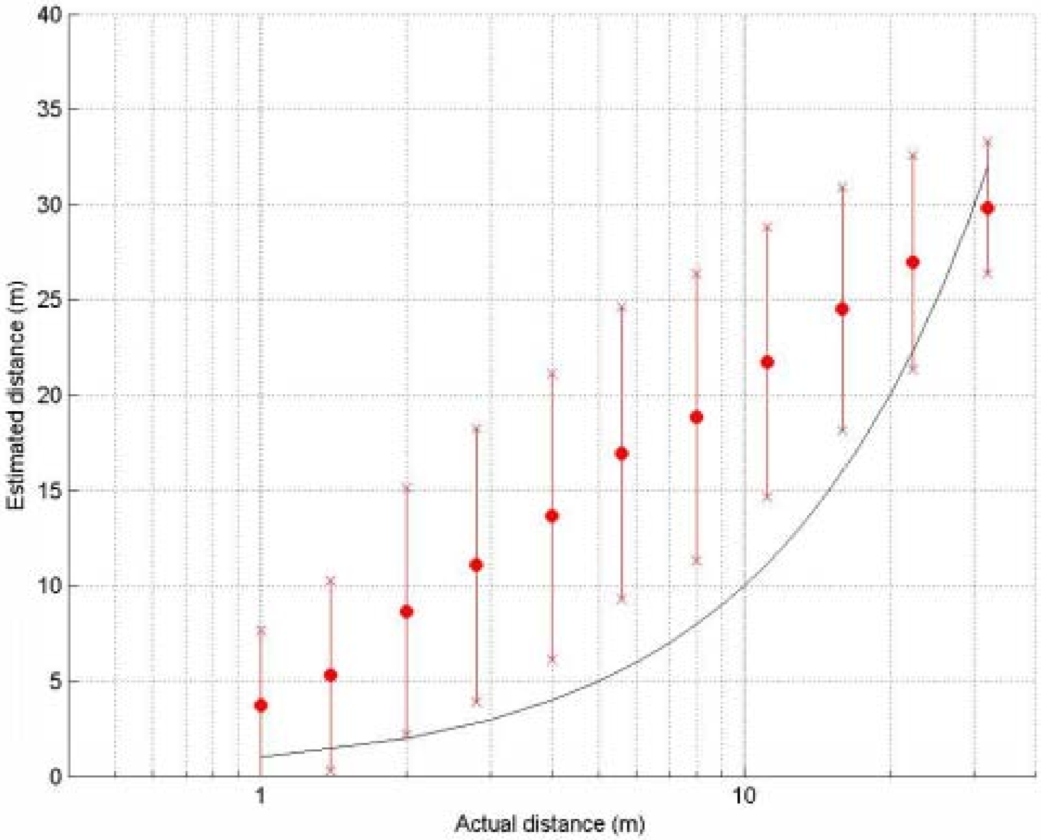


FIG. 7. Estimated vs. actual TLD (m) from the 2-party experiment for reference tokens. Circles are means, bounded lines show +/- 1 s.d., and black curve shows 1:1 correspondence.

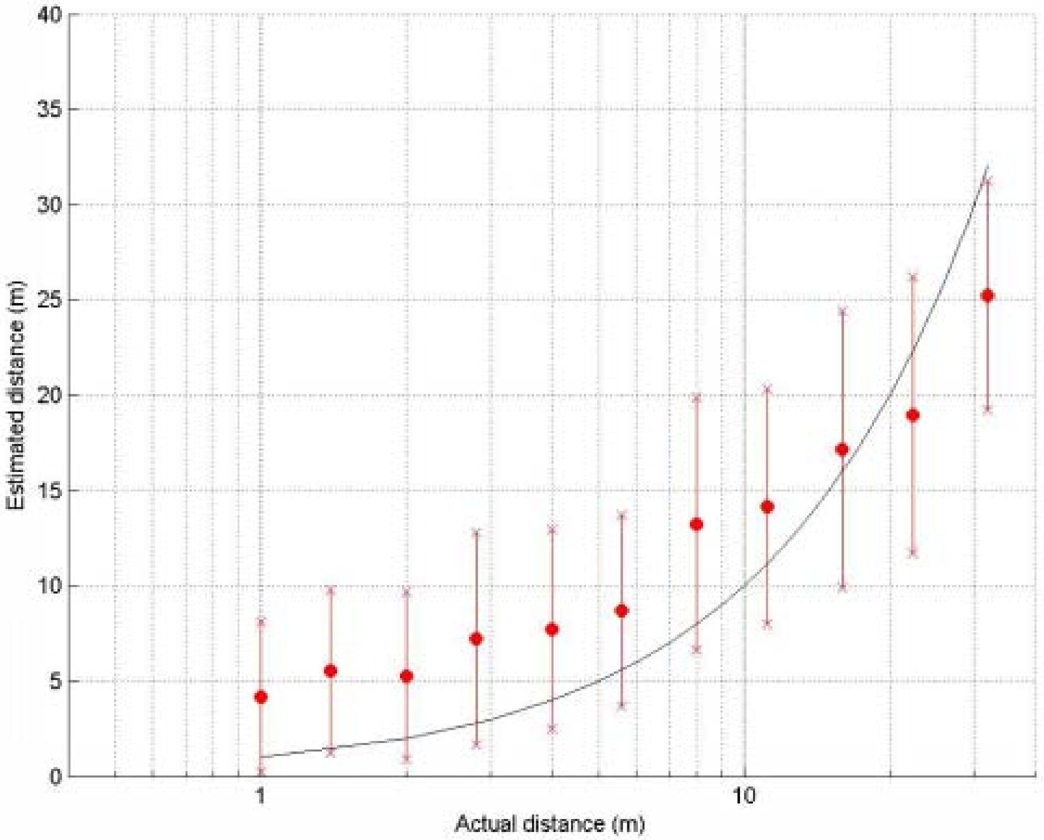


FIG. 8. Estimated vs. actual TLD (m) from the 3-party experiment for reference tokens. Circles are means, bounded lines show +/- 1 s.d., and black curve shows 1:1 correspondence.

FIGS. 7 and 8 show the averaged results for the reference recording stimuli, across all listeners and talkers for the two-party and three-party distance estimates, respectively. The filled circles show the mean values at each actual distance, and the lines bounded by crosses show plus/minus one standard deviation. The general trend shows increasing estimated distance with increasing actual distance. For the two-party case, note that on average the estimated distance exceeds the actual distance by as much as a factor of four (e.g., at 1m), except for the 32m actual distance. Aside from the endpoints, the standard deviation (range 5.0-7.7m) remains relatively constant but is rather large throughout the range. For the three-party case, there is less of a tendency for overestimating distance, and the standard deviation is similar to the 2-party case (range 4.3-7.3m, excepting endpoints). However, there is more confusion among adjacent distances, especially in the 1m-2m range and the 2.8m-5.6m range. This suggests that vocal effort changes are not significant within these ranges. The relatively small acoustic changes noted in the above Tables 2-5 for these ranges also provide evidence that vocal effort changes are insignificant within these ranges.

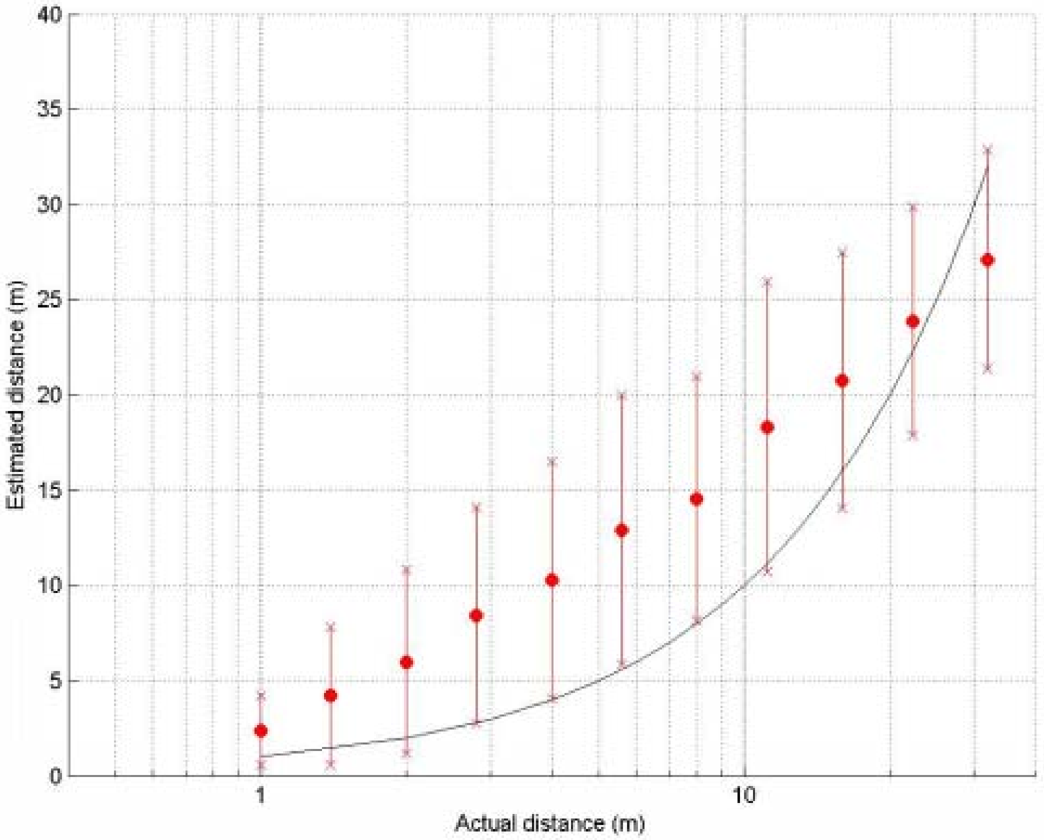


FIG. 9. Estimated vs. actual TLD (m) for the 2-party experiment for HLsyn tokens. Circles are means, bounded lines show +/- 1 s.d., and black curve shows 1:1 correspondence.

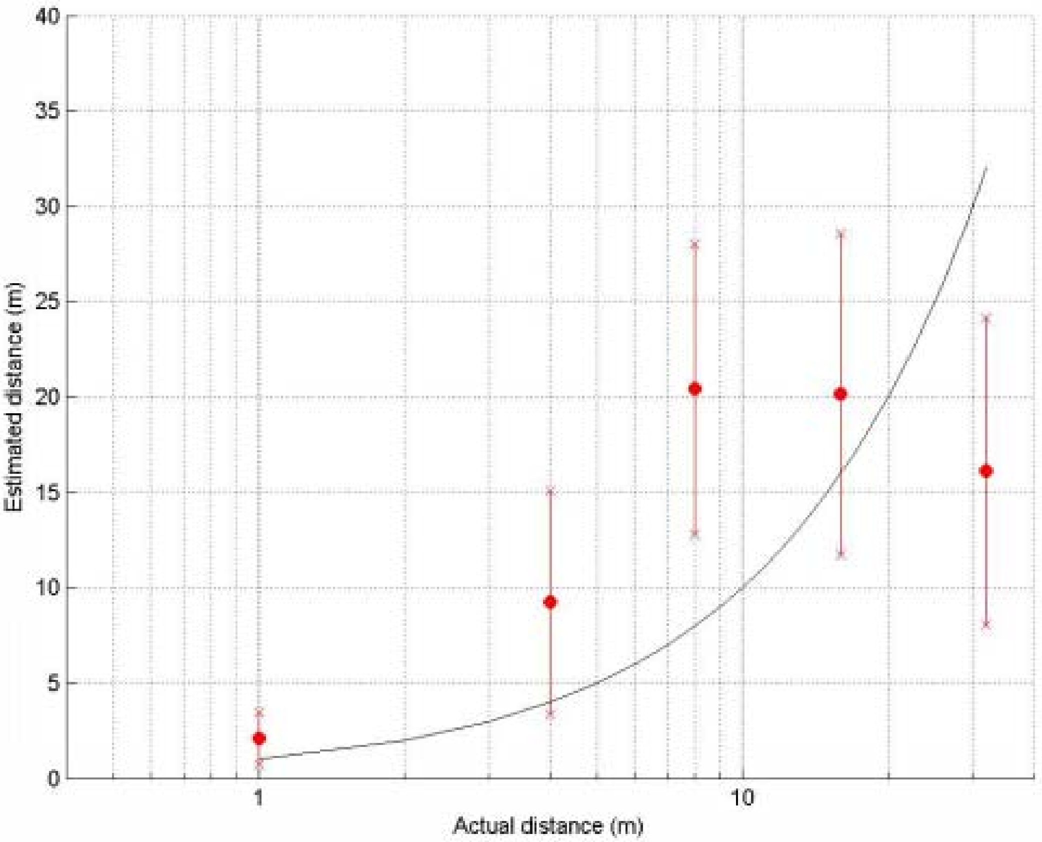


FIG. 10. Estimated vs. actual TLD (m) for the 2-party experiment for MELP-modified tokens. Circles are means, bounded lines show +/- 1 s.d., and black curve shows 1:1 correspondence.

FIG. 9 shows the averaged results for the HLsyn stimuli that vary in f0 only for the two-party distance estimates. Again, the general trend shows increasing estimated distance with increasing actual distance. Note the similarity of the HLsyn token results shown in FIG. 9 and the reference recording results of FIG. 7, suggesting that attenuation due to TLD and changes in f0 are significant cues for perception of TLD, an observation supported by the work of Brungart et al. and Brungart & Scott. On average the estimated distance exceeds the actual distance by as much as a factor of two (e.g., at 1m), except for the 32m actual distance. Initial results of one listener's perception of TLD for the MELP modified speech are shown in FIGS. 10 and 11 for the 2-party and 3-party cases, respectively. The stimuli for these tests were generated by passing a subset of reference tokens spoken at 1m TLD through the modified MELP vocoder. The vocoder altered the tokens' f0 and overall intensity (what Brungart calls the "production level") to mimic the changes that occur for TLDs of 4m, 8m, 16m, and 32m. The 1m tokens were obtained by passing the reference tokens through the vocoder without modifying them. The responses in FIG. 6 demonstrate the feasibility of the system for the TLDs of 1m, 4m, and 8m in that the mean estimated distance increases with the actual distance, although overestimation of the distances is similar to what is seen with the reference tokens and HLsyn tokens in FIGS. 7 and 9. The response confusion among the 8m, 16m and 32m distances illustrate the limitations of the current MELP vocoder algorithm that must be addressed. One such limitation is a maximum f0 of 200 Hz, which prevented many of the tokens from having f0 modified appropriately for the desired TLD, especially for the female voices. For several of the included tokens, the high f0 resulted in the vocoder output sounding unnatural. The insufficient change in f0 and the unnaturalness of some tokens likely contributed to the response confusion among the larger TLDs.

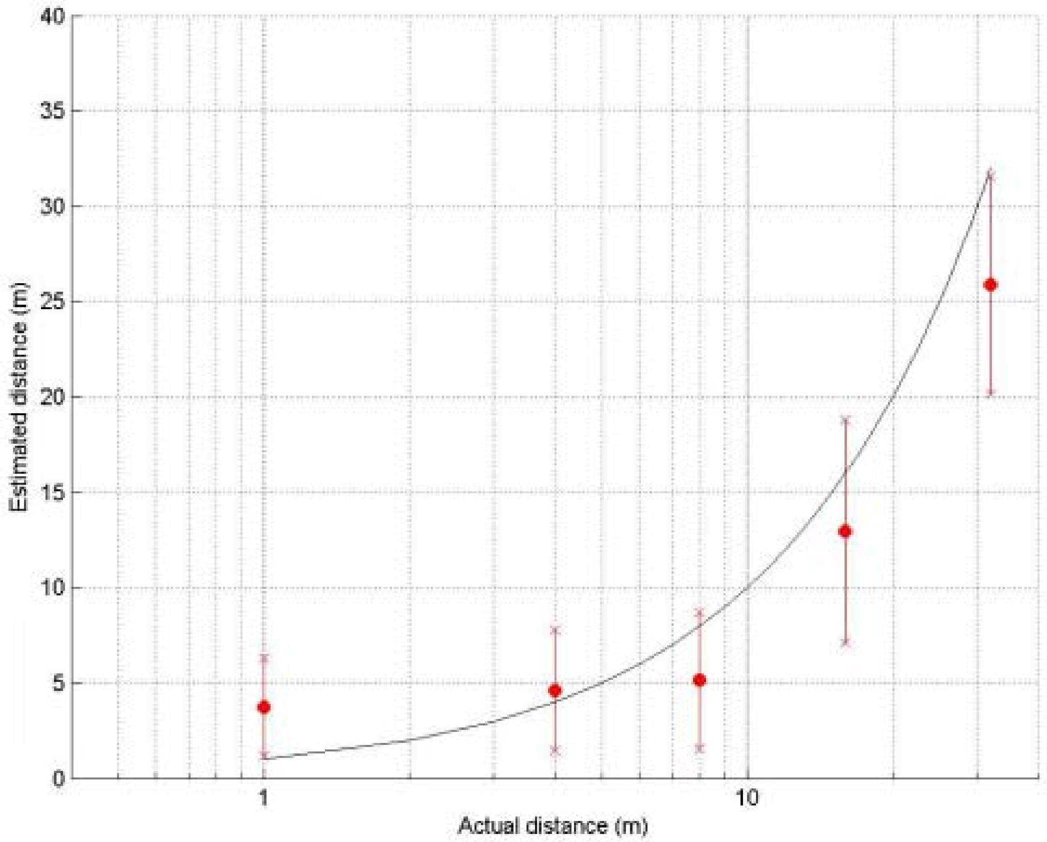


FIG 11. Estimated vs. actual TLD (m) from the 3-party experiment for MELP-modified tokens. Circles are means, bounded lines show +/- 1 s.d., and black curve shows 1:1 correspondence.

FIG. 11 shows the 3-party results from the same listener whose results are shown in FIG. 10. Note that here the response confusion appears reversed relative to that in FIG. 10: there is confusion among the 1m, 4m, and 8m TLDs but distinction among the 8m, 16m, and 32m TLDs. This suggests that (for this listener at least) although the vocal effort changes in f0 may not have been realistic for the tokens at the larger TLDs, in the absence of attenuation due to TLD the MELP modified vocal effort changes were sufficient for discrimination. Thus even without the range of f0 change needed to mimic the actual changes in vocal effort, these initial results support the feasibility of the approach. The results in FIG. 6 and 7 together suggest that the listener relied on the attenuation due to TLD for estimating distances less than 8m, and vocal effort changes for estimating distances greater than 8m.

IV. CONCLUSIONS

The reference recordings obtained during this work represent the largest corpus of real speech that varies in vocal effort as it relates to TLD. The analysis of these recordings for changes in f0, intensity, and formant amplitudes as a function of TLD provides an initial look at what aspects of speech vary systematically with vocal effort, as vocal effort is modified to compensate for changes in TLD. Perceptual testing of synthetic and modified speech tokens revealed that in both the two-party and three-party experimental paradigms, listeners tend to overestimate TLD, but to a lesser extent when only the vocal effort cues are available.

Algorithms for changing vocal effort, and thus perceived TLD, would have numerous military and commercial applications. Potential military applications include (a) virtual training environments with enhanced acoustic cues for consistency with visual cues of nearby or distant talkers; (b) command-and-control operations using virtual audio displays where distance perception may enhance the separation of talkers when added to directional cues; (c) remote battlefield environments, such as several operators remotely controlling unmanned aircraft or SWORDS robots, where a virtual acoustic space incorporating distance and direction cues is necessary to maintain awareness of their machines' relative locations when the operators' physical separation from each other does not mimic that of the machines; and (d) an aid for unfamiliar route navigation, in which a virtual voice can speak to the person(s) navigating the route (e.g., "Over here!") and the voice contains perceptual cues relating to the listener-to-target distance as well as direction. A related potential application of an algorithm to modify vocal effort would be to simulate a change in emotion, or specifically urgency. For a command-and-control operation, a command issued as a shout nearby a listener would naturally carry more urgency or import to the listener than the same command spoken conversationally.

Potential commercial applications include (a) enhanced realism in virtual audio spaces, such as motion pictures and video games, to provide auditory cues for talker-to-listener distance that are consistent with visual cues, (b) providing speaker separation in teleconferencing on a monaural channel; and (c) aids for the visually impaired, in which a virtual voice can be perceived at the location of a distant object of interest. Lastly, these algorithms could suggest solutions or partial solutions to the inverse problem: given a speech signal picked up at a listener’s location, what is the distance to the talker? This distance-estimation problem may also have potential military and commercial applications (e.g., remote machine awareness of its distance to talkers).

V. FUTURE WORK

Several potential manipulations in speech that occur with TLD need further investigation. First, segmental duration may increase with TLD, and may vary with the type of segment. For example, vowel durations may increase more rapidly than non-resonant consonant durations. Second, increases in f0 may not be simply a shift across the entire voiced segment, but may involve a larger range of f0 – perhaps starting and ending at similar f0 values but increasing the excursion during the word. Third, changes in formant amplitudes may also not be constant across the vocalic portions of the words. Fourth, the vocal tract resonances during fricatives may change with TLD. Fifth, the intensity of the non-resonant phonemes (e.g., fricatives) may change differently from the resonant phonemes (e.g., vowels) with TLD. While not an exhaustive list of the potential changes that may occur with speech over TLD, these five suggestions are the most probable next steps for investigation.

Work with the MELP vocoder algorithm revealed its limitations for this application. First, the bandwidth of the MELP is restricted to 4 KHz, which limits the quality of the modified speech particularly for female speakers. Manipulation of the formant amplitudes to mimic those that occur naturally requires a method whereby all target formant amplitude values can simultaneously be altered through multivariable optimization, to obtain both the required formant amplitude and spectral tilt changes with TLD.

VII. PUBLICATIONS

[1] Kalgaonkar, K.; Clements, M. *Vocal tract area based formant tracking using particle filter* ICASSP 2008. March 31 2008-April 4 2008 Page(s):3405 - 3408

[2] Kalgaonkar, K.; Clements, M. *Vocal Tract Area Function with both Lip and Glottal Losses.* The 9th Interspeech Antwerp Belgium Aug 2007.

[3] Harold Cheyne, Kaustubh Kalgaonkar, Mark Clements & Patrick Zurek, *Talker-to-Listener Distance Effects on speech Production and Perception.* American Speech Language Hearing Association Conf, Novemeber 2007 Boston MA USA.

[4] Harold Cheyne, Kaustubh Kalgaonkar, Mark Clements & Patrick Zurek, *Talker-to-Listener Distance Effects on Speech Production.* The Journal of Acoustical Society of America (Accepted for publication).

VIII. REFERENCES

1. Bungart, D.S., Scott, K.R. (**2001**). “The effects of production and presentation level on the auditory distance perception of speech,” J. Acoust. Soc. Am. **110**, 425-440.

2. Brungart, D.S., Kordik, A.J., Das, K., Shaw, A.K. (**2002**). “The effects of f0 manipulation on the perceived distance of speech,” Proceedings of the International Conference on Spoken Language Processing (Denver, Colorado), 1641-1644.

3. Liénard, J-S., Di Benedetto, M-G. (**1999**). “Effect of vocal effort on spectral properties of vowels,” J. Acoust. Soc. Am. **106**, 411-422.

4. Eriksson, A., Traunmüller, H. (**2002**). “Perception of vocal effort and distance from the speaker on the basis of vowel utterances,” Percept. Psychophys. **64**, 131-139.

5. Traunmüller, H., Eriksson, A. (**2000**). “Acoustic effects of variation in vocal effort by men, women, and children,” J. Acoust. Soc. Am. **107**, 3438-3451.

6. Tassa, A., Liénard, J.S. (**2000**). “A New Approach to the Evaluation of Vocal Effort by the PSOLA Method,” The European Student Journal of Language and Speech **00.01** (online only at http://www.essex.ac.uk/web-sls/papers/00-01/00-01.html).

7. Bond, Z.S., Moore, T.J., Gable, B. (**1989**). “Acoustic-phonetic characteristics of speech produced in noise and while wearing an oxygen mask,” J. Acoust. Soc. Am. **85**, 907-912.

8. Summers, W. Van, Pisoni, D.B., Bernacki, R.H., Pedlow, R.I., Stokes, M.A. (**1988**). “Effects of noise on speech production: Acoustic and perceptual analyses,” J. Acoust. Soc. Am. **84**, 917-928.

9. Huber, J.E., Stathopoulos, E.T., Curione, G.M., Ash, T.A., Johnson, K. (**1999**). “Formants of children, women, and men: The effects of vocal intensity variation,” J. Acoust. Soc. Am. **106**, 1532-1542.

10. Rostolland, D. (**1982**). “Acoustic features of shouted voice,” Acoustica **50**, 118-125.

11. Hanson, H.M., Chuang, E.S. (**1999**). “Glottal characteristics of male speakers: Acoustic correlates and comparison with female data,” J. Acoust. Soc. Am. **106**, 1064-1077.

12. Hanson, H.M., Stevens, K.N.(**2002**). “A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using Hlsyn,” J. Acoust. Soc. Am. **112,** 1158-1182.

13. Stevens, K.N., Bickley, C.A. (**1991**). “Constraints among parameters simplify control of Klatt formant synthesizer,” J. Phon. **19**, 161-174.

14. Klatt, D.H., Klatt, L.C. (**1990**). “Analysis, synthesis, and perception of voice quality variations among female and male talkers,” J. Acoust. Soc. Am. **87**, 820-857.

15. Brungart, D.S. (**2000**). “A Speech-Based Auditory Display,” Proceedings of the AES 109th Convention (Los Angeles, California), 1-11.

16. Barnwell, T.P. III, George, E.B., McCree, A.V., Truong, K.K., Viswanathan, V.R. (**1996**) "A 2.4 kbit/s MELP Coder Candidate for the New U.S. Federal Standard," Proceedings of the International Conference on Acoustics, Speech, and Signal Processing..

17. Proakis, J.G., Manolakis, D.G. (**1996**) *Digital Signal Processing: Principles, Algorithms and Applications* (Prentice-Hall, Englewood Cliffs, NJ).

18. Oppenheim, R.W., Schafer, A.V. (**1989**) *Discrete-Time Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ).

19. Morris, R., Clements, M. (**2002**). “Modification of Formants in the Line Spectrum Domain,” IEEE Signal Processing Letters **9**, 19-21.