Formant Tracking Using Context-Dependent Phonemic Information

Minkyu Lee, Jan van Santen, Bernd Möbius, and Joseph Olive

***Abstract*—A new formant-tracking algorithm using phoneme information is proposed. Conventional formant-tracking algorithms obtain formant tracks by analyzing the acoustic speech signal using continuity constraints without any additional information. The formant-tracking error rate of the conventional methods is reportedly in the range of 10%–20%. In this paper, we show that if text or phoneme transcription of speech utterances is available, the error rate can be significantly reduced. The basic idea behind this approach is that given the phoneme identity, formant-tracking algorithms can have a better clue of where to look for formants. The algorithm consists of three phases: 1) analysis, 2) segmentation and alignment, and 3) formant tracking by the Viterbi searching algorithm. In the analysis phase, formant candidates are obtained for each analysis frame by solving the linear prediction polynomial. In the segmentation and alignment phase, the text corresponding to the input speech utterance is converted into a sequence of phoneme symbols. Then, the phoneme sequence is time aligned with the speech utterance. A hidden Markov model (HMM) based automatic segmentation algorithm is used for forced-time alignment. For each phoneme segment, nominal formant frequencies are assigned at the center of each phoneme segment. Then nominal formant tracks for the entire utterance are obtained by interpolating the nominal formant frequencies. In order to compensate for the coarticulation effect, different interpolation methods are used depending on the phonemic context. The interpolation process makes the formant-tracking algorithm robust to possible segmentation errors made by the HMM-based segmentation algorithm. As a result, the proposed formant-tracking algorithm does not require highly accurate alignment/segmentation. Finally, a set of formants is chosen from the formant candidates in such a way that the resulting formant tracks come close to the nominal formant tracks while satisfying the continuity constraints. The algorithm is tested using natural speech utterances and the performance is compared against formant tracks obtained by the conventional method using continuity constraints only. The new algorithm significantly reduces the formant-tracking error rate (5.03% for male and 3.73% for female) over the conventional formant-tracking algorithm (13.00% for male and 15.82% for female).**

***Index Terms*—Automatic segmentation, coarticulation, dynamic programming, formant tracking, speech analysis.**

Manuscript received January 23, 2002; revised May 31, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shrinkanth Narayanan.

M. Lee is with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: minkyul@research.bell-labs.com).

J. van Santen is with the School of Science and Engineering, Oregon Graduate Institute of Science and Technology, Oregon Health and Science University, Beaverton, OR 97006 USA (e-mail: vansanten@cslu.ogi.edu).

B. Möbius is with the IMS, University of Stuttgart, D-70174 Stuttgart, Germany (e-mail: moebius@ims.uni-stuttgart.de; bernd.moebius@ims.uni-stuttgart.de).

J. Olive is with DARPA/IPTO, Arlington, VA 22203 USA (e-mail: joseph.olive@darpa.mil).

Digital Object Identifier 10.1109/TSA.2005.851904

I. Introduction

FORMANT tracking is an important speech analysis problem that can benefit many speech applications such as speech recognition, compression, and synthesis. It is also a useful tool for phoneticians studying the mechanisms of human speech production. In the Bell Laboratories' Text-To-Speech (TTS) system [1], formants are used for selecting concatenation units from recorded speech. Since only a limited number of acoustic units is stored in the inventory and speech is synthesized by concatenating the units, it is important to be able to choose the best candidate for each synthesis unit (diphone, triphone, etc). Formants can also be used for checking unit compatibility to determine whether or not any two synthesis units are connectable in terms of spectral discrepancy [1]. Thus, reliable formant tracking is one of the crucial components in the construction of the Bell Laboratories' TTS system. The recent trend in TTS development is to increase the size of acoustic inventory. When a large amount of speech material has to be processed, it would be prohibitive to rely on human intervention to correct errors made by automatic formant-tracking methods.

For decades, researchers have worked to improve the performance of speech formant-tracking algorithms. Nevertheless, state-of-the-art formant-tracking algorithms are not reliable enough for unsupervised, automatic usage. Even though the errors can be obvious to the human eyes when displayed in a longer time frame, a human might not do much better than the automatic formant trackers given only local information. This observation has led to methods that impose continuity constraints on the formant selection process [2], [3]. However, errors tend to occur when the continuity constraints are either too strong or too weak. Finding the proper strength of the continuity constraints is not a trivial task. In highly transient phone boundaries, such as consonant–vowel transitions, the continuity constraints often cause tracking errors [4]–[6].

In this paper, we propose a new algorithm for tracking speech formant trajectories. In some speech applications, text transcription of the speech utterances is available. Phoneme transcription can be generated automatically from the texts. It is also possible to time align the phoneme transcription with the acoustic speech signal using various signal processing techniques. For a given speech interval, the identity of a phoneme can be determined and nominal formant values for the phoneme are also available from the literature [7]. In the proposed method, the nominal formant values are used as references for formant searching. We describe how much improvement can be achieved by using phoneme information in addition to the conventional constraints such as the continuity constraints. Performance is compared with previous

work [8], [9] in which formant tracking is performed without text or phoneme information.

II. Algorithm

The formant-tracking algorithm consists of three phases: 1) analysis, 2) segmentation/alignment, and 3) formant track selection, as shown in Fig. 1. In the analysis phase, formant candidates are obtained by solving the linear prediction polynomial from LPC analysis on pre-emphasized speech. In the segmentation phase, input text is converted into a sequence of phoneme symbols, and the phoneme sequence is time aligned with the acoustic speech utterance. Finally, in the formant-tracking phase, the best set of formant frequencies is selected from the candidates, based on minimum cost criteria. For each analysis frame, we choose a set of formant candidates that is closest to the nominal formant tracks while satisfying the continuity constraints. In this section, each block will be explained in more detail.

A. Speech Analysis

First, autocorrelation LPC analysis is performed on pre-emphasized input speech. An LPC order of 12 is used for speech data collected at the sampling rate of 11.025 kHz. Thus, ten complex poles (five conjugate pole pairs) will be used to model five formants and the extra two poles are for the spectral tilt that might not have been compensated for by the pre-emphasis process. Pitch-asynchronous LPC coefficients are calculated every 5 ms. A Hamming window of 25 ms is applied to each analysis frame. Formant frequency candidates are obtained by root solving the prediction polynomial using Bairstow's method [10]. Only complex roots are considered as candidates for final formants. If a complex root pair is

$$s, s^* = -\sigma \pm j2\pi F. \tag{1}$$

Then, F is the formant frequency and the corresponding bandwidth is approximately 2σ [11]. Therefore, there are, at most, six formant candidates for each analysis frame from twelfth-order LPC analysis. The goal of the proposed algorithm is to select smooth formant tracks from the candidate set.

B. Text-to-Phoneme Transcription

In the text-to-phoneme conversion module, the text transcription of input speech is converted into a sequence of phoneme symbols. We use the text analysis front-end of the Bell Laboratories' TTS system [1] for text-to-phoneme conversion. The front-end includes components such as sentence-boundary detection, abbreviation expansion, number expansion and so on. Then morphological analysis is performed for lemmatization of inflected words using a finite-state machine. Finally, the words are converted into phoneme sequences using dictionary lookup and letter-to-sound rules.

A probabilistic system is also used to generate alternative pronunciations for a given phoneme sequence produced by TTS's front-end. This is required in order to cope with the possible mismatches between the TTS phoneme sequence and actual speech produced by the speaker. During the automatic segmentation stage, hidden Markov model (HMM) based algorithms attempt to determine which alternative pronunciation is actually produced by the speaker.

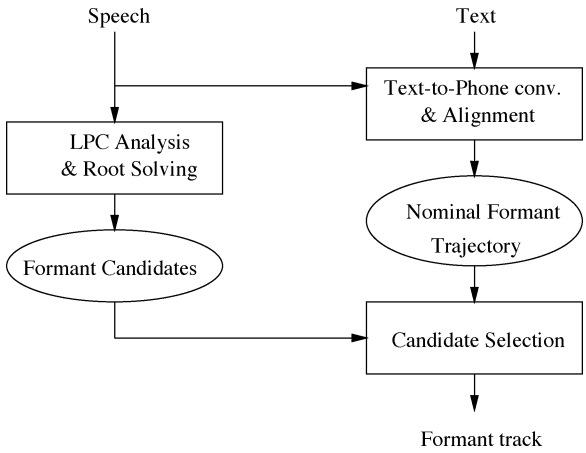

Fig. 1. Overall system block diagram for formant tracking.

C. Automatic Speech Segmentation/Alignment

The next step is to segment the acoustic signal into phoneme segments and align them with the phoneme sequence. This task is often called the forced-alignment problem. Although manual segmentation and alignment is considered to be the most accurate method [12], the segmentation results from human experts do not always agree [13]–[15]. In addition, manual segmentation can be time consuming. Many algorithms have been proposed in order to obtain a consistent automatic formant tracker. Using HMM-based speech recognizers combined with the Viterbi search is the most popular and natural method [16], [17]. Ho [18] used phoneme-dependent HMM-based formant tracking for voice morphing. Van Santen and Sproat [19] applied edge detectors in different frequency bands. The boundaries are then combined with lowest-cost path algorithms applied to finite-state transducers. Hosom [20] incorporated acoustic-phonetic information such as voicing, glottalization, and burst-related impulses, into a phonetic alignment system.

The performance of automatic forced-alignment algorithms is often compared against the performance of human experts. For speaker-independent methods, state-of-the-art forced-alignment algorithms report about 86% accuracy within 20 ms and about 96% accuracy within 40 ms for clean speech [17], [21]. Speaker-dependent methods can achieve slightly higher accuracy.

In this study, we implement a conventional HMM-based speech segmentation algorithm with Viterbi search. Improving the performance of automatic segmentation algorithms is beyond the scope of this study. Rather, we will show that using state-of-the-art automatic segmentation algorithms is sufficient for improving the performance of conventional formant-tracking algorithms. A schematic diagram of a typical HMM-based segmentation system is shown in Fig. 2. Speech parameters are extracted from the input speech. For HMM training, we use a set of HMM triphone models trained for Bell Laboratories' speaker independent automatic speech recognition (ASR) engine [22]. The HMM model is trained using *Wall Street Journal* corpora. Each triphone is modeled using a

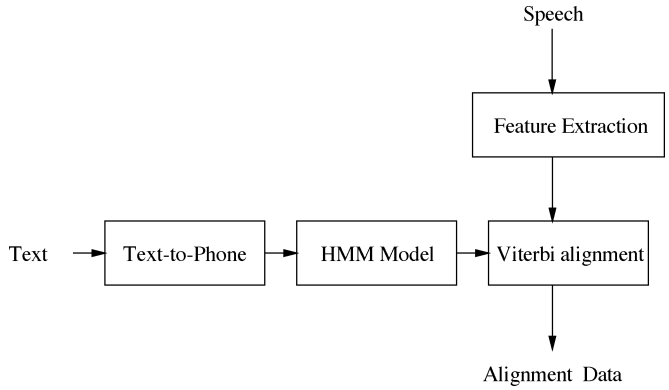

Fig. 2. Block diagram of an HMM based segmentation/alignment system.

six-state left-to-right HMM topology. Speech analysis frame length is 25 ms and the analysis window is shifted by 5 ms. The topology does not allow skipping state. Therefore, the minimum triphone length that the topology can model is 30 ms. Input phoneme sequence $P = (p_1, p_2, p_3, \ldots, p_N)$ can be easily converted into a sequence of triphones, $T = (t_1, t_2, t_3, \ldots, t_N)$. Then, using the HMM triphone models, an HMM network λ for input text can be constructed by sequencing the initial state of a triphone model p_{i+1} with the last state of the preceding triphone model p_i.

In the next stage for automatic speech segmentation, an observation vector $O = (o_1, o_2, o_3, \ldots, o_T)$ is extracted every 5 ms from input speech. For the observation sequence O and the underlying HMM model λ, the best state sequence $q = (q_1, q_2, q_3, \ldots, q_T)$ can be found using the Viterbi time-alignment algorithm [23].

In order to evaluate the performance of the automatic segmentation algorithm, the segmentation results are compared with hand-segmented phone boundaries. A total of 4,223 utterances (2049 male and 2174 female speech) are segmented both manually and automatically. The total number of phone boundaries are 45 217 and 18 597 for male and female speech, respectively. Table I shows the performance of the segmentation algorithm using the manual segmentation results as references. About 81.6% for female and 77.7% for male speech have segmentation errors of less than 20 ms. And about 95.5% for female and 92.5% for male less than 40 ms. About 98.9% of female and 96.7% of female speech have segmentation errors less than 100 ms. The average absolute differences are 12.28 and 13.23 ms for female and male speech, respectively. Table II summarizes the results for different phoneme class pairs. The summary is for both male and female utterances. Phonemes are classified into seven classes: vowel, nasal, fricative, affricate, stop, liquid, and glide. One can observe that vowel–vowel, vowel–liquid, stop–stop, and fric–fric transitions show relatively larger segmentation error. Not surprisingly, those groups are the classes that human expert labelers do not always agree.

D. Nominal Formant Tracks

This section explains how to obtain nominal formant tracks using the phoneme boundaries. Given phoneme identity and location, nominal formant tracks are constructed using nominal formant values for each phoneme. Tables III and IV show nominal formant values for U.S. English male and female speakers, respectively. The values are the averages of three U.S. male speakers measured by a human expert. For female speech, formant values 15% higher are used. The results in Section III will show that this approximation for female formant values is appropriate for our purpose.

Voicing probabilities of 1.0 for voiced, 0.0 for unvoiced and stops, and 0.6 for nasal sounds are used. The voicing probability acts as a confidence measure for the nominal formant frequencies, i.e., the nominal formant values are highly credible for voiced sounds, but not for mixed and unvoiced sounds. Nominal formant values and a voicing probability are assigned to the temporal center of each phoneme segment. Nominal formants and a voicing probability for the frames between the center points are interpolated for smooth nominal formant trajectories. In order to account for the coarticulation effect, various interpolation methods are used depending on the phonemic context. Linear interpolation is applied in general. But, for some special phoneme sequences where strong coarticulation may occur, nonlinear interpolation is adopted. It is well known that some phoneme classes affect the sound of surrounding phonemes more than others [24]. In order to consider the contextual effects on formant frequencies, we classify voiced sounds into subgroups whose sounds have similar coarticulation effects. The subgroups are 1) liquids (/r/ and /l/), 2) glides (/w/ and /y/), 3) nasals (/m/, /n/, and /ŋ/), and 4) the remaining vowels. When phonemes from different subgroups are connected, formants of one or both phonemes can be changed dramatically due to the coarticulation effect. In this section, such classes are explained in further detail. Only voiced sounds are considered for contextual effects because formants are not well defined for unvoiced sounds.

1) Transitions Between Vowels and Glides (or Liquids): In English, a vowel followed by a glide between word or syllable boundaries forms a diphthong. However, in vowel–glide (or glide–vowel) transitions at syllable or word boundaries, the glide functions as an independent phoneme. The influence of the vowel and glide sounds on each other is mainly on the length of the vowel part of the transition. A vowel–glide transition is more abrupt and the glide sounds have a longer duration [7]. In a glide–vowel transition, conversely, the spectrogram looks like a mirror image of the vowel–glide transition. Liquid sounds are generally very resistant to coarticulation. Therefore, the influence of vowel and liquid sounds on each other is similar to what we found in the transition between a vowel and a glide.

Nonuniform phoneme length change results in a nonlinear formant transition near the phoneme boundary. In order to account for this effect, nonlinear interpolation is performed on nominal formant values. An interpolation factor for a phoneme I to J transition is described as

$$a_t = \left(\frac{t - I_{\text{mid}}}{J_{\text{mid}} - I_{\text{mid}}} \right)^2, \quad I_{\text{mid}} \le t \le J_{\text{mid}} \tag{2}$$

or

$$a_t = \frac{-1 \cdot (t - J_{\text{mid}})^2}{(I_{\text{mid}} - J_{\text{mid}})^2} + 1, \quad I_{\text{mid}} \le t \le J_{\text{mid}} \tag{3}$$

TABLE I
SUMMARY OF AUTOMATIC SEGMENTATION ACCURACY

Gender	Abs. Mean	< 5 msec	< 10 msec	< 20 msec	< 40 msec	< 100 msec
female	12.28 msec	29.6 %	54.7 %	81.6 %	95.5 %	98.9 %
male	13.23 msec	26.9 %	49.9 %	77.7 %	92.5 %	96.7 %

TABLE II
SUMMARY OF SEGMENTATION ACCURACY FOR MOST FREQUENT PHONEME CLASS TRANSITIONS. TEST DATABASE INCLUDES BOTH MALE AND FEMALE UTTERANCES. TRANSITIONS WITH MORE THAN 500 PAIRS IN THE TEST DATABASE ARE LISTED. VOWEL–VOWEL, VOWEL–LIQUID, STOP–STOP, AND FRIC–FRIC TRANSITIONS SHOW RELATIVELY LARGER SEGMENTATION ERROR

Category	Total	Abs. Mean Diff.(ms)	<5 ms (%)	<10 ms (%)	<20 ms (%)	<40 ms (%)	<100 ms (%)
stop-vowel	7906	11.66	64.62	80.74	95.62	99.29	99.81
vowel-stop	7314	13.75	68.29	79.89	93.74	98.95	99.95
fric-vowel	5501	10.23	45.96	69.22	94.36	99.69	99.82
vowel-fric	5406	11.57	58.14	77.80	94.14	99.30	99.85
vowel-nasal	4447	13.14	43.06	59.41	84.82	97.68	99.62
liquid-vowel	4387	17.81	86.92	92.30	96.83	98.54	99.73
vowel-liquid	4041	24.03	49.00	57.46	71.76	88.25	98.69
vowel-vowel	2410	34.16	55.56	61.95	70.71	81.78	96.76
nasal-vowel	2269	9.77	78.05	91.49	98.90	99.38	99.82
glide-vowel	1705	16.88	71.38	80.12	90.91	97.65	99.82
stop-liquid	1690	14.20	31.66	49.05	81.60	98.22	99.47
fric-stop	1635	12.52	64.28	76.15	93.39	97.80	99.82
stop-fric	1343	13.20	70.14	79.23	90.25	97.54	100.00
nasal-stop	1332	12.57	62.99	75.75	90.47	97.75	99.77
nasal-fric	879	16.63	61.77	70.76	89.19	99.43	100.00
liquid-stop	815	13.69	82.45	91.53	96.81	99.14	100.00
glot-vowel	722	11.33	71.61	84.35	94.46	99.03	100.00
liquid-fric	675	8.39	81.19	91.85	99.26	99.70	100.00
stop-stop	667	23.30	65.52	71.21	80.81	93.85	100.00
stop-glide	547	19.06	29.80	39.85	63.07	94.88	100.00
fric-fric	510	20.23	60.39	70.78	84.71	95.88	99.61

where a_t is an interpolation factor at time t, and I_{mid} and J_{mid} are the center points of phoneme I and J, respectively. Equation (2) is used when the formants of phoneme I needs to be extended into the region of phoneme J, i.e., progressive assimilation. Equation (3) is used for regressive assimilation, where the formant of a preceding phoneme is corrupted by the following phoneme. The nonlinear interpolation factor is depicted in Fig. 3.

The ith nominal formant at time t is calculated as

$$F_t^i = (1.0 - a_t) \cdot FI_n^i + a_t \cdot FJ_n^i, \quad i = 1 \ldots 4 \qquad (4)$$

where F_t^i is the ith nominal formant interpolated at time t, and FI_n^i and FJ_n^i are the ith nominal formant for phonemes I and J, respectively.

Examples of nonlinear interpolation for transitions between vowels and glides are illustrated in Fig. 4(a). Using nonlinear interpolation, the resulting nominal formant tracks can better guide the formant selection algorithm to true formant tracks.

2) Vowels to and From Nasals: Nasal sounds are produced when the vocal tract is constricted at some point in the oral cavity with the velum lowered. The air from the glottis flows through the nasal tract. The mouth traps acoustic energy at some frequencies resulting in antiresonance in the spectrum. The nasal sounds typically show broad formants at low frequency and no prominent formants at mid-to-high frequency range.

Due to the special characteristics of nasal sounds, the spectrogram of the transition region shows very clear and abrupt discontinuities at the boundaries between vowels and nasals. Therefore, the continuity constraints must be relaxed to allow abrupt formant changes near nasals as depicted in Fig. 4(b).

3) Vowels to and From Fricatives and Stops: Fricative sounds are produced by turbulence at a constriction in the vocal tract. Since the source of the sound is at the constriction, not at the glottis, the acoustics of the sound is less influenced by

TABLE III
VOICING PROBABILITY AND NOMINAL FORMANT VALUES FOR AVERAGE U.S. MALE SPEAKERS

IPA Symbol	Class	Example	P_v	F_1	F_2	F_3	F_4
ŋ	Nasal	sing	0.6	236	1978	2660	3660
m	Nasal	met	0.6	215	1338	2380	3380
n	Nasal	net	0.6	227	1636	2534	3534
ə	Schwa	about	1	509	1564	2536	3536
eI	Diphthong	bait	1	408	2063	2583	3583
aI	Diphthong	buy	1	596	1646	2515	3515
oʊ	Diphthong	boat	1	423	1054	2390	3390
aʊ	Diphthong	down	1	694	1246	2538	3538
ɔI	Diphthong	boy	1	418	1254	2371	3371
a	Vowel	bat	1	674	1703	2495	3495
ɛ	Vowel	bet	1	579	1703	2538	3538
I	Vowel	bit	1	381	1868	2610	3610
ɑ	Vowel	Bob	1	679	1177	2483	3483
ʊ	Vowel	book	1	387	1255	2433	3433
ɔ	Vowel	bought	1	633	1070	2501	3501
ʌ	Vowel	but	1	602	1336	2451	3451
ɝ	Vowel	bird	1	413	1317	1634	2634
u	Vowel	boot	1	301	1146	2381	3381
i	Vowel	beat	1	279	2324	2678	3678
l	Liquid	let,tell	1	339	894	2634	3634
r	Liquid	rent	1	431	1213	1720	2720
w	Glide	wit	1	265	563	2395	3395
y	Glide	you	1	255	2258	3123	4123

the vocal tract resonance. Even with alveolar (e.g., /s/ and /z/) and palato–alveolar (e.g., /ʃ/ and /ʒ/) fricatives, which are produced further back in the oral cavity, the front oral cavity is considerably smaller than the back cavity. Consequently, fricatives do not display formant structure. For similar reasons, stop sounds display little formant structure.

At the transitions of vowels to and from fricatives (or stops), the formant values approach those of the adjacent vowels, especially for voiced fricatives and stops. As shown in Fig. 4(a), we extend the nominal formants of the neighboring vowels into fricative or stop regions.

The reason we define nominal formant tracks for fricatives and stops, although they do not have well-defined formant structure, is that, in conventional formant-tracking algorithm, we observe many formant errors in the voiced sound that follows fricatives or stops. Most of the errors are due to the continuity constraints that is trying to connect spurious formants found in fricatives region with the formant candidates in the following voiced sound.

E. Formant Tracking

The next step is to choose the best set of formants from the candidates that is close to the nominal formant tracks and that satisfy the continuity constraints. This can be achieved by means of dynamic programming.

The problem is to choose N formants from n candidates over K analysis frames, where K is the total number of frames in the utterance. For 11.025-kHz speech, there are, at most, six formant candidates, and we are interested in four formant tracks. Thus, here, we have $N = 4$ and $n = 6$. At each frame k there are L_k ways to map (assign) the candidates to formants. The L_k mappings can be identified as

$$L_k = \binom{n}{N} = \frac{n!}{(n-N)!N!} \tag{5}$$

where n is the number of candidates obtained in the previous analysis phase and N is the desired number of formants.

The formants are chosen from the candidates based on minimal total cost criteria, which is calculated from several cost functions: local cost, frequency change cost, and transition cost. The local cost λ_{kl} of the lth mapping at the kth frame is based on the nth formant bandwidth B_{kln} and the deviation of the nth formant frequency F_{kln} from nominal formant frequencies for the phoneme Fn_n

$$\lambda_{kl} = \sum_{n=1}^{N} \nu_n \left\{ \beta_n B_{kln} + \mu_n \frac{|F_{kln} - Fn_n|}{Fn_n} \right\} \tag{6}$$

where ν_n is the voicing probability, β_n determines the cost of bandwidth broadening for the nth formant, and μ_n determines the cost of deviations from the nominal frequency of the nth formant. The $\mathrm{n}(=6)$ candidate formant frequencies F_{kln} and bandwidths B_{kln} are calculated by root solving the twelfth-order LPC polynomial from the LPC analysis block shown in Fig. 1. Given a complex root pair (1), the formant frequency is F and the corresponding bandwidth is approximately 2σ [11].

The frequency change cost ξ_{kljn} between the lth mapping at frame k and the jth mapping at frame $k-1$ for the nth formant is defined as

$$\xi_{kljn} = \left\{ \frac{F_{kln} - F_{k-1jn}}{F_{kln} + F_{k-1jn}} \right\}^2. \tag{7}$$

The quadratic cost function is to penalize any abrupt formant frequency change across analysis frames. Using (7), a transition cost δ_{klj} can be defined as a weighted sum of the frequency change cost of each individual formant

$$\delta_{klj} = \psi_k \sum_{n=1}^{N} \alpha_n \xi_{kljn} \tag{8}$$

where α_n determines the relative cost of interframe frequency changes in the nth formant. The term ψ_k is designed to modulate the weight of the formant continuity constraints based on the acoustic/phonetic context of the frames. For example, formant trajectories are often discontinuous across silence–vowel, vowel–consonant, and consonant–vowel boundaries. One should avoid placing the continuity constraints on those boundaries. The ψ_k can be any kind of similarity measures or inverse of distance measures such as interframe spectral distance measures in the LPC or cepstral domain. We use a simple stationarity measure based on the signal rms energy, by which the weight of the continuity constraints can be reduced when rms energy changes abruptly. It is defined in terms of the relative signal rms difference between the current and previous frames

$$\psi_k = -log\left(\frac{|\mathrm{rms}_k - \mathrm{rms}_{k-1}|}{\mathrm{rms}_{\mathrm{MAX}}} \right) \tag{9}$$

TABLE IV
VOICING PROBABILITY AND NOMINAL FORMANT VALUES FOR AVERAGE U.S. FEMALE SPEAKERS

IPA Symbol	Class	Example	P_v	F_1	F_2	F_3	F_4
ŋ	Nasal	sing	0.6	271.4	2274.7	3059.0	4209.0
m	Nasal	met	0.6	247.2	1538.7	2737.0	3887.0
n	Nasal	net	0.6	261.0	1881.4	2914.1	4064.1
ə	Schwa	about	1	585.3	1798.6	2916.4	4066.4
eI	Diphthong	bait	1	469.2	2372.4	2970.4	4120.4
aI	Diphthong	buy	1	685.4	1892.9	2892.2	4042.2
oʊ	Diphthong	boat	1	486.4	1212.1	2748.5	3898.5
aʊ	Diphthong	down	1	798.1	1432.9	2918.7	4068.7
ɔI	Diphthong	boy	1	480.7	1442.1	2726.6	3876.6
a	Vowel	bat	1	775.1	1958.4	2869.2	4019.2
ɛ	Vowel	bet	1	665.8	1958.4	2918.7	4068.7
I	Vowel	bit	1	438.1	2148.2	3001.5	4151.5
ɑ	Vowel	Bob	1	780.8	1353.5	2855.4	4005.4
ʊ	Vowel	book	1	445.0	1443.2	2797.9	3947.9
ɔ	Vowel	bought	1	727.9	1230.5	2876.1	4026.1
ʌ	Vowel	but	1	692.3	1536.4	2818.6	3968.6
ɝ	Vowel	bird	1	474.9	1514.5	1879.1	3029.1
u	Vowel	boot	1	346.1	1317.9	2738.1	3888.1
i	Vowel	beat	1	320.8	2672.6	3079.7	4229.7
l	Liquid	let,tell	1	389.8	1028.1	3029.1	4179.1
r	Liquid	rent	1	495.6	1394.9	1978.0	3128.0
w	Glide	wit	1	304.8	647.4	2754.2	3904.2
y	Glide	you	1	293.2	2596.7	3591.4	4741.4

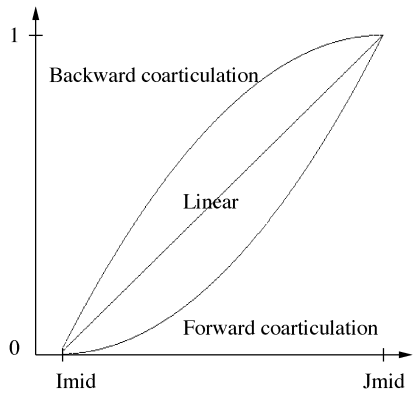

Fig. 3. Interpolation factor.

with rms_k being the signal rms in the kth analysis frame and $\mathrm{rms}_{\mathrm{MAX}}$ is the maximum rms energy in the utterance. The operation $|x|$ is the absolute value of x. When there is an abrupt rms energy change, as in the case of silence–vowel, vowel–consonant, and consonant–vowel boundaries, ψ_k becomes small reducing the effect of the frequency change cost ξ_{kljn}. Obviously, this stationarity measure is too simple to detect all possible phone boundaries. The proposed idea of utilizing phone identity and its nominal formant frequencies (6) is to prevent forced restriction across the phone boundary.

Finally, the minimum total cost of choosing candidate formant frequencies over K analysis frames with L_k mappings at each frame can be defined as

$$C = \sum_{k=1}^{K} \min_{l \epsilon L_k} D_{kl}. \tag{10}$$

As shown in Fig. 5, the mapping cost D_{kl} for the lth mapping at the kth frame is obtained from

$$D_{kl} = \lambda_{kl} + \min_{j \epsilon L_{k-1}} \gamma_{klj} \tag{11}$$

where λ_{kl} is given in (6) and γ_{klj}, the connection cost from the jth mapping at frame $k-1$ to the lth mapping in frame k, is defined by the recursion

$$\gamma_{klj} = \delta_{klj} + D_{k-1j}. \tag{12}$$

In the present implementation, the constants α_n, β_n, and μ_n are independent of n. The values of α_n and β_n are determined empirically [8], while the value of μ_n is varied to select the optimal weight for the cost of deviation from the nominal formant frequencies.

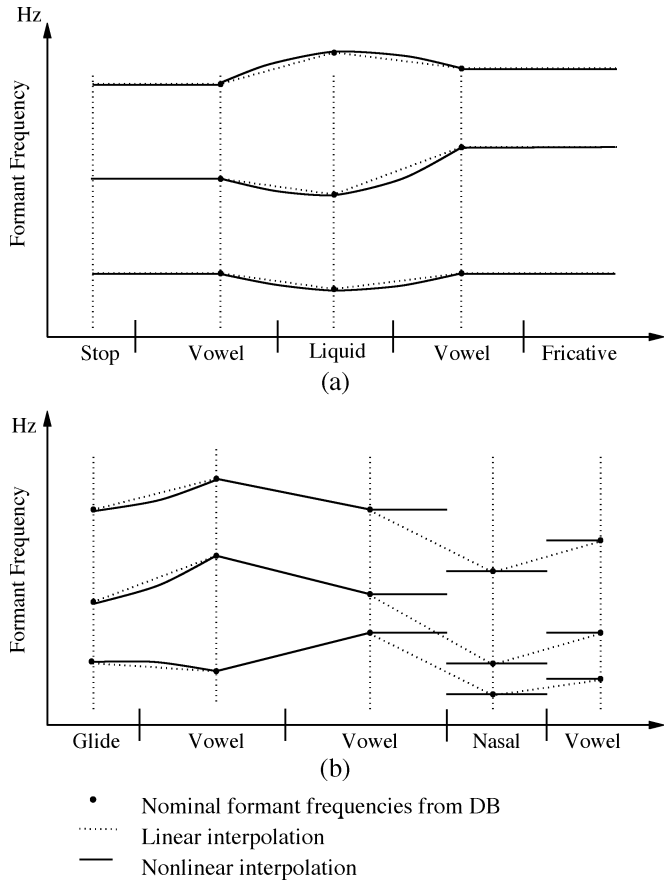

Fig. 4. Nominal formant tracks obtained by linear and nonlinear interpolation.

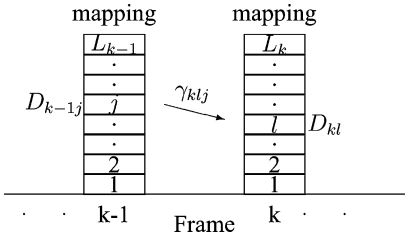

Fig. 5. Mapping cost D_{kl} is the sum of local cost λ_{kl} and the minimum connection cost γ_{klj}. The γ_{klj} is calculated using the frequency change cost δ_{klj} and the mapping cost D_{k-1j} of the previous frame.

III. Results

A total of 600 American English utterances from two male and two female speakers are tested (e.g., 150 utterances from each speaker). The utterances are recorded at 44.1 kHz, downsampled to 11.025 kHz and stored in 16-bit integer. Table V summarizes the average number of words and the average length of the utterances for each speaker.

The performance of the proposed algorithm is compared with that of a formant tracker using only the continuity constraints. The results are visually inspected by overlaying the formant tracks on top of the corresponding spectrogram. Formant-tracking error rate is calculated for voiced sounds only, because formants are not well-defined for unvoiced sounds.

Formant-tracking errors are identified based on the following rules. If a tracker misses the first formant and consequently assigns the second to the first formant and the third to the second formant, the algorithm is considered to have made errors in all three formants. As such, if it detects the first formant but misses the second formant, hence assigning the third to the second formant, the second and third formant are counted as errors. Accordingly, the number of errors tends to increase with the formant number. If the first and third formants are correctly identified while the second formant is placed at the wrong frequency, only the second formant is labeled as an error.

Table VI lists the number of phones with formant-tracking errors for different methods for each gender group. The first three rows P1, P2, and P3 are for the newly proposed algorithm with different weightings μ_n on the cost function (6). The last row, denoted as CC, shows the results for the conventional formant tracker using the continuity constraints only, i.e., μ_n is set to zero [8]. Smaller μ_n means less cost for deviation from the nominal formant values, resulting in relatively stronger continuity constraints. In other words, P1 has weak continuity constraints producing formant tracks that are heavily influenced by the nominal formant values. On the other hand, CC has strong continuity constraints resulting in smoother formant tracks. Total errors are the number of phones that have one or more formant errors (any formant error regardless of F_1, F_2, F_3, or F_4). The next four columns, F_1–F_4 errors, show how formant errors are distributed over the formant number. Since a formant error can occur at both F_1 and F_2, the first four formant errors do not add up to the total errors.

As would be expected, the new proposed algorithm presents much improved results (5.03% for male and 3.73% for female speech) over the conventional formant tracker CC (13.00% for male and 15.82% for female). Among the three tests with different μ_n values, P3 ($\mu_n = 4$) shows the best performance. Notice that for the conventional method a large portion of the errors are at F_1 ($228/824 = 27.67\%$ for male and $182/815 = 22.33\%$ for female) and F_2 ($479/824 = 58.13\%$ for male and $412/815 = 50.55\%$ for female). On the other hand, the new proposed algorithm P3 has errors occurring in the F_2 or F_3 track over 90% of the male and female speech. In the unit selection process for the Bell Laboratories' TTS system, F_1 and F_2 are more heavily weighted in the acoustic unit selection process than F_3. In the acoustic unit selection process, mismatches of the first formant are more critical than mismatches of the higher formants.

In general, segmentation errors of 40 ms or less often do not critically affect the performance of the proposed formant-tracking algorithm because the nominal formant values near phoneme boundaries are smoothly interpolated (except nasals). In order to verify this postulate, we also tested the proposed algorithm using manually segmented phone boundaries. Table VII shows the performance of formant-tracking errors using manually segmented phoneme boundaries. The error rates are 4.67% and 1.20% for male and female test utterances, respectively. Compared to this result, formant tracking based on automatic segmentation produces a slightly worse performance (5.03% for male and 3.73% for female speech). Nonetheless, it is a great improvement from the conventional formant-tracking method.

Figs. 7 and 8 show formant-tracking results for a speech sample shown in Fig. 6 using the conventional method and the proposed method, respectively. The conventional method

TABLE V
SUMMARY OF THE FORMANT TRACKING TEST SET. EACH SPEAKER HAS 150 UTTERANCES. ALL NUMBERS ARE IN AVERAGE PER UTTERANCE. VOWEL–LIKE PHONEMES ARE LISTED IN TABLE III

SpeakerID	Gender	Length (sec)	# Words	# Vowel-like phones
JAC	Male	4.54	10.48	23.96
KBB	Male	2.89	8.28	18.30
MER	Female	5.92	7.73	15.55
KER	Female	3.68	8.77	18.80

TABLE VI
SUMMARY OF FORMANT TRACKING ERROR USING AUTOMATIC SPEECH ALIGNMENT ALGORITHM (P1–P3) AND THE CONVENTIONAL FORMANT TRACKING METHOD (CC). FOR METHODS P1, P2, AND P3, THE VALUES OF μ_n ARE SET TO 10, 7, AND 4, RESPECTIVELY. FOR CC METHOD, μ_n IS SET TO ZERO TO IMPOSE THE CONTINUITY CONSTRAINTS ONLY. THUS, P1 HAS THE STRONGEST DEPENDENCY ON THE NOMINAL FREQUENCIES WHILE CC HAS THE STRONGEST CONTINUITY CONSTRAINT

Gender (# phones)	Method	Total Errs(%)	F_1 errs	F_2 errs	F_3 errs	F_4 errs	μ_n
Male (6340)	P1	370 (5.83 %)	24	157	278	351	10
	P2	322 (5.08 %)	13	128	231	310	7
	P3	319 (5.03 %)	13	131	239	303	4
	CC	824 (13.00 %)	228	479	721	811	0
Female (5153)	P1	217 (4.21 %)	20	78	105	198	10
	P2	196 (3.80 %)	17	71	99	176	7
	P3	192 (3.73 %)	15	59	91	175	4
	CC	815 (15.82 %)	182	412	601	787	0

TABLE VII
SUMMARY OF FORMANT TRACKING ERROR USING MANUAL SPEECH ALIGNMENT. PHONE BOUNDARIES ARE IDENTIFIED BY HUMAN EXPERT LABELERS $\mu_n = 4$

Gender(# phones)	Total Errs(%)	F_1 errs	F_2 errs	F_3 errs	F_4 errs
Male (6340)	296 (4.67 %)	14	125	214	292
Female (5153)	62 (1.20 %)	10	26	50	53

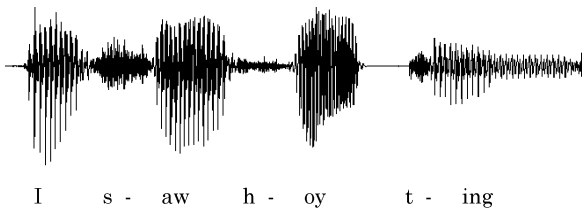

Fig. 6. Speech waveform in "I saw hoyting guys."

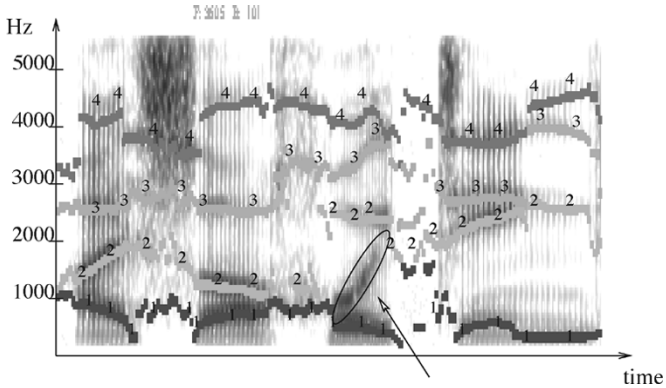

Fig. 7. Spectrogram and formant tracks—conventional method.

CC (Fig. 7) clearly misses the second formant track near the diphthong indicated by an arrow. This is probably because the continuity constraints forced the tracking algorithm to render the second formant in the "oy" segment continuous with the second formant of the previous voiceless fricative "h" near 2400 Hz. This is a typical error occurring when the continuity constraints place too much emphasis on connecting the formant tracts of the vowel segment to the preceding fricative segment. Fig. 8 shows the correct tracking results obtained by the proposed method, The new algorithm found the second formant near the nominal formant values at about 1250 Hz of the "oy" segment.

Errors still tend to occur when there is strong coarticulation. Figs. 9–11 show an example of formant-tracking error which is due to coarticulation. According to Table III, the first three nominal formant values for a vowel /ɑ/ are around 674, 1703, and 2495 Hz while formant values for a retroflex sound /r/ are around 431, 1203, 1720 Hz. When the two phonemes are connected as /ɑ/-/r/, the formant tracks in the early part of /ɑ/

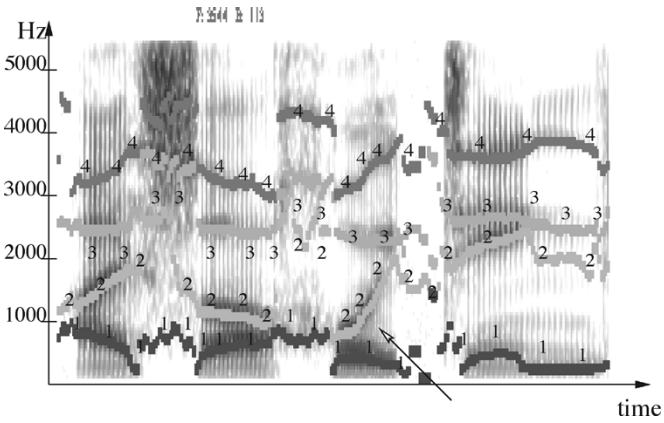

Fig. 8. Spectrogram and formant tracks—proposed method.

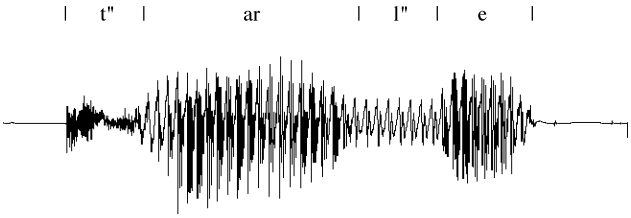

Fig. 9. Speech waveform—/tar le/in "See the tar letting guys."

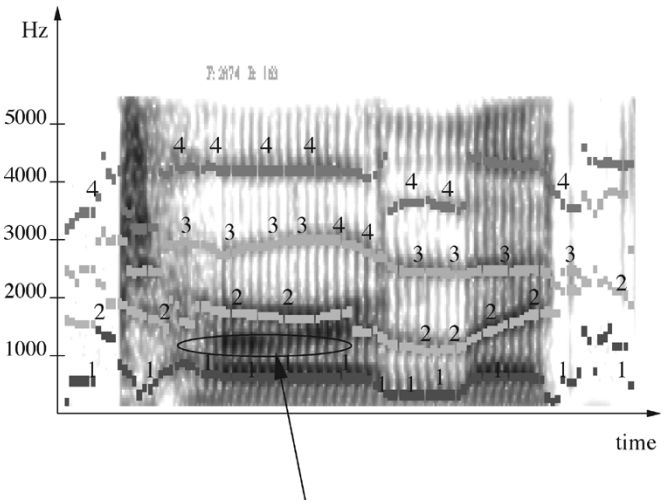

Fig. 10. Spectrogram and formant tracks—conventional method.

often show the second formant around 1200 Hz, which is where the second formant of /r/ is located. The proposed method successfully selects the right formants for the /r/ part, while the conventional method fails. However, both methods fail to detect the low second formant of the /ɑ/ part which is influenced by the following /r/ sound. In this case, the nominal formant values in the /ɑ/ segment are invalid due to heavy coarticulation.

To summarize, the above results indicate that if a text or transcription of an input utterance is available, the formant-tracking performance can be improved considerably.

IV. Summary

We have presented a new formant-tracking algorithm using the knowledge of phoneme identity of the analysis frame. The algorithm consists of three phases: 1) LPC analysis and root solving, 2) segmentation and alignment using HMM-based forced-alignment algorithm, and 3) formant tracking by the Viterbi searching algorithm. The proposed method searches for formant tracks that are close to the nominal formant tracks while satisfying the continuity constraints. The algorithm is tested using natural speech utterances and the performance is compared against formant tracks obtained by conventional method, which is using the continuity constraints only. The new algorithm significantly reduces the formant-tracking error rate (5.03% for male and 3.73% for female) over a formant-tracking algorithm using only the continuity constraints (13.00% for male and 15.82% for female).

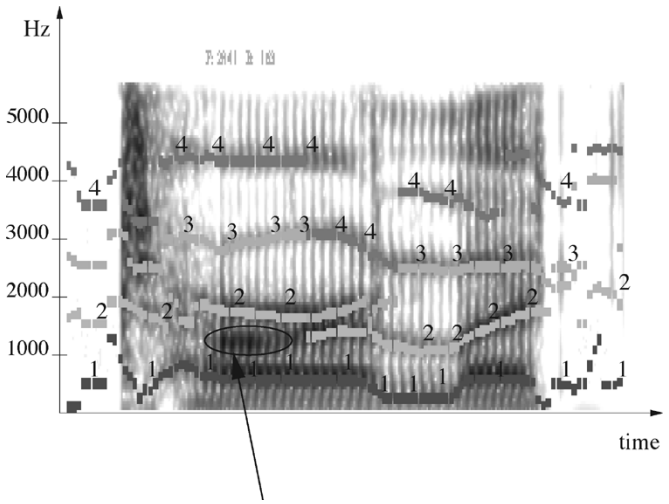

Fig. 11. Spectrogram and formant tracks—Proposed method.

References

[1] R. Sproat, Ed., *Multilingual Text-To-Speech Synthesis: The Bell Labs Approach*. Boston, MA: Kluwer, 1998.
[2] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 57, no. 2, pp. 634–648, 1970.
[3] S. McCandless, "Automatic formant extraction using linear prediction," *J. Acoust. Soc. Amer.*, vol. 54, no. 1, p. 339, 1974.
[4] M. Hunt, "A robust formant-based speech spectrum comparison measure," in *Proc. IEEE Int. Conf. Acoustics and Speech Signal Processing*, 1985, pp. 1117–1120.
[5] G. E. Kopec, "Formant tracking using hidden markov models and vector quantization," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-34, no. Aug., pp. 709–729, 1986.
[6] S. S. Seneff, "An auditory-based speech recognition strategy: application to speaker-independent vowel recognition," presented at the Speech Recognition Workshop, 1986.
[7] J. P. Olive, A. Greenwood, and J. Coleman, *The Acoustics of American English Speech: A Dynamic Approach*. New York: Springer, 1993.
[8] D. Talkin, "Speech formant trajectory estimation using dynamic programming with modulated transition costs," AT&T Bell Laboratories, Tech. Rep. 11 222-870 720-07TM, 1987.
[9] M. Lee, J. van Santen, B. Möbius, and J. Olive, "Formant tracking using segmental phonemic information," in *Proc. Eur. Conf. Speech Communication and Technology (Eurospeech)*, vol. 6, Budapest, Hungary, 1999, pp. 2789–2792.
[10] R. W. Hamming, *Numerical Methods for Scientists and Engineers*. New York: McGraw-Hill, 1962.
[11] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Murray Hill, NJ: Bell Laboratories, 1978.
[12] Z. Wu, C. Barras, E. Geoffrois, and M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Commun.*, vol. 33, pp. 5–22, 2001.

[13] P. Cosi, D. Falavigna, and M. Omologo, "A preliminary statistical evaluation of manual and automatic segmentation discrepancies," in *Proc. Eur. Conf. Speech Communication and Technology*, vol. 5, 1991, pp. 1947–1950.
[14] A. Ljolje, J. Hirschberg, and J. P. H. van Santen, "Automatic speech segmentation for concatenative inventory selection," in *Progress in Speech Synthesis*, J. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, Eds. New York: Springer, 1997, pp. 305–312.
[15] R. Cole, B. T. Oshika, M. Noel, T. Lander, and M. Fanty, "Labeler agreement in phonetic labeling of continuous speech," in *Proc. Int. Conf. Spoken Language Processing*, Yokohama, Japan, 1994, pp. 2131–2134.
[16] C. W. Wightman and D. Talkin, "The aligner:text-to-speech aligment using markov models," in *Progress in Speech Synthesis*, J. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, Eds. New York: Springer, 1997, pp. 313–324.
[17] B. L. Pellom and J. H. L. Hansen, "Automatic segmentation of speech recorded in unknown noisy channel characteristics," *Speech Commun.*, vol. 25, pp. 97–116, 1998.
[18] C.-H. Ho, D. Rentzos, and S. Vaseghi, "Formant model estimation and transformation for voice morphing," in *Proc. Int. Conf. Spoken Language Processing*, Denver, CO, 2002, pp. 2149–2152.
[19] J. van Santen and R. Sproat, "High-accuracy automatic segmentation," in *Proc. Eur. Conf. Speech Communication and Technology*, vol. 6, 1999, pp. 2809–2812.
[20] J.-P. Hosom, "Automatic time alignment of phonemes using acoustic-phonetic information," Ph.D. dissertation, Graduate Inst. Sci. Technol., Oregon Health Sci. Univ., Beaverton, 2000.
[21] F. Malfrere, O. Deroo, and T. Dutoit, "Phonetic alignment: speech synthesis based vs. hybrid hmm/ann," in *Proc. Int. Conf. Spoken Language Processing*, Sydney, Australia, 1998, pp. 1571–1574.
[22] W. Reichl and W. Chou, "Robust decision tree state tying for continuous speech recognition," *IEEE Proc. Speech Audio Processing*, vol. 8, no. 5, pp. 555–566, Sep. 2000.
[23] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[24] P. Ladefoged, *A Course in Phonetics*, 3rd ed. Orlando, FL: Harcourt Brace Jovanovich, 1993.

Minkyu Lee received the B.S. degree in electrical engineering from Seoul National University, Seoul, Korea, in 1986, the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology in 1988, and the Ph.D. degree from the Department of Electric Engineering, University of Florida, Gainesville, in 1996.

He is a Member of Technical Staff at Bell Laboratories, Lucent Technologies, Murray Hill, NJ. He was a Research Scientist with the Electronics and Telecommunications Research Institute (ETRI), Korea, from 1988 to 1992 and a Voice Modeling Engineer at Voxware, Inc., Princeton, NJ, from 1996 to 1998. His research interest includes multilingual TTS synthesis, speech coding, recognition, and voice-over-IP signal processing.

Jan van Santen received the Ph.D. degree in mathematical psychology from the University of Michigan, Ann Arbor, in 1979.

He is the Director of the Center for Spoken Language Understanding and a Professor of biomedical engineering and computer science at the Oregon Health and Science University, Beaverton. His current responsibilities consist of teaching, conducting research, and managing the Center for Spoken Language Understanding. His research focuses on mathematical modeling of prosody, signal processing, and computational linguistics. A key growing focus of his work and of the Center is on basic and applied speech and language technology research for communication disorders. He was a Member of Technical Staff, Bell Laboratories, Lucent Technologies, Murray Hill, NJ, from 1984 to 2000 and an Associate Research Scientist, New York University, from 1981 to 1984.

Bernd Möbius received the Ph.D. degree from the University of Bonn, Bonn, Germany, in 1992.

He is an Associate Professor of phonetics at the Institute of Natural Language Processing (IMS), University of Stuttgart, Stuttgart, Germany. His research is in linguistics and phonetics, including speech production and perception, prosody, phonology, and some aspects of computational linguistics. He is currently focusing on prosody in the context of speech production and on unit selection speech synthesis. Before joining the IMS in 1999, he was a Research Scientist at Bell Laboratories, Lucent Technologies, Murray Hill, NJ, from 1993 to 1998.

Joseph Olive received the degree in physics and the M.A. degree in music composition from the University of Chicago, Chicago, IL.

He has over 30 years of experience in research and development at Bell Laboratories, Lucent Technologies, Murray Hill, NJ, and 19 years of experience in management. He has been the world leader in the research of TTS synthesis and has managed a world-class team in computer dialogue systems and human-computer communication. In his role as Director of Speech Research and CTO of Lucent's Business Unit Lucent Speech Solutions, he supervised the productization of Bell Laboratories' core speech technologies: ASR, TTS, and speaker verification. He also led the dialogue research team to create a "next-generation" dialogue system for e-mail reading and navigation. As a Research Manager, he has been instrumental in developing the careers of some of the members in his group by guiding their research and supervising numerous postdoctorates and summer students. While he was a graduate student, his research consisted of computational atomic physics requiring the intensive use of computers for the computation of electrons distribution functions. He was also a member of the University of Chicago's computer center. After leaving the University of Chicago, he combined his interest in computation and music and began research in acoustics and signal processing. He used the M.A. degree he received in music composition to pursue a side career in writing music for small chamber groups, orchestras, computer music, and an opera for a computer, soprano, and a small ensemble.

Dr. Olive was a recipient of the National Endowment for the Arts grant in 1974 to write a computer opera. He was also the recipient of the Bell Laboratories' Distinguished Member of Technical Staff award in 1984.