6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

FORMANT TRACKING USING SEGMENTAL PHONEMIC INFORMATION

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

Bell Labs, Lucent Technologies, 600 Mountain Avenue, Murray Hill, NJ 07974, USA
*IMS, University of Stuttgart, Azenbergstrasse 12, D-70174 Stuttgart, Germany

10.21437/Eurospeech.1999-615

ABSTRACT

A new formant tracking algorithm using phoneme dependent nominal formant values is tested. The algorithm consists of three phases: (1) analysis, (2) segmentation, and (3) formant tracking. In the analysis phase, formant candidates are obtained by solving for the roots of the linear prediction polynomial. In the segmentation phase, the input text is converted into a sequence of phonemic symbols. Then the sequence is time aligned with the speech utterance. Finally, a set of formant candidates that are close to the nominal formant estimates while satisfying the continuity constraints are chosen. The new algorithm significantly reduces the formant tracking error rate (3.62%) over a formant tracking algorithm using only continuity constraints (13.04%). We will also discuss how to further reduce the tracking error rate.

INTRODUCTION

In the Bell Labs' Text-To-Speech (TTS) system [1], a limited number of acoustic units is stored in the inventory table. Therefore, it is important to be able to choose the best candidate for each synthesis unit (diphone, triphone, etc). Formants values can be used for selecting the best units as well as for testing unit compatibility to determine whether any two synthesis units are connectable in term of spectral discrepancy [1]. Thus, reliable formant tracking can be one of the crucial components in TTS system construction, where a huge amount of speech data has to be processed. Due to the size of the speech corpus, it would be prohibitive to rely on human intervention for formant tracking error correction.

For decades, researchers have put efforts into improving 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 are obvious to the human eye 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, they still tend to generate errors by enforcing the continuity constraints too strongly or too weakly. Especially in highly transient phone boundaries such as consonant-vowel transitions, continuity constraints often cause tracking errors [4],[5],[6] .

Fortunately, in the TTS system construction process, transcriptions of the speech utterances are available. During speech corpus recording, a speaker is asked to read a set of texts that are carefully selected. From the text, the phonemic transcription can be generated automatically. Then, the transcription can be time aligned with the acoustic speech signal using signal processing techniques. Using this forced time alignment, the exact time stamp for each phonemic event can be obtained.

In this paper, we test a new algorithm for tracking speech formant trajectories using segmental phonemic information. Given a speech interval, it is assumed that the phonemic identity and nominal formant values for the phoneme are available. This assumption holds always in TTS applications. The implementation is based on previous work [7] in which only continuity constraints were used. We will show how much improvement can be achieved by using phonemic information for formant tracking.

ALGORITHM

The formant tracking algorithm consists of three phases: (1) analysis, (2) segmentation/alignment, and (3) formant track selection. In the analysis phase, formant candidates are obtained by LPC analysis on pre-emphasized speech. Formant candidates are obtained by solving for the roots of the linear prediction polynomial. In the segmentation phase, the input text is converted into a sequence of phonemic symbols, and the phonemic symbols are time aligned with the speech utterance. Finally, in the formant tracking phase, the best combination 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 are closest to the nominal formant estimates while satisfying the continuity constraints.

Speech Analysis

Autocorrelation LPC analysis is performed on the pre-emphasized speech. An LPC order of 12 is used for speech data collected at a sampling rate of 11.025 kHz. Thus, ten complex poles (five conjugate pairs) will be used to model five formants and the extra two poles for the spectral tilt that might have not been compensated for by the pre-emphasis process. Pitch-asynchronous LPC coefficients are calculated every 5 ms. A Hamming window of 25ms is applied to each analysis frame. Formant frequency candidates are calculated by solving the prediction polynomial using Bairstow's method [8]. Only complex poles are considered as formant candidates.

Text to Phonetic Transcript

Given the input text, a sequence of graphemes is converted into a sequence of phonemic symbols. We have used the text analysis front-end of the Bell Labs TTS system [1]. The front-end includes components such as sentence-boundary detection, abbreviation expansion, number expansion, etc. 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 that is not part of the TTS system is used to generate alternative pronunciations for a given phoneme sequence produced by TTS's front-end. This is required because of possible mismatches between the TTS phoneme sequence and actual speech.

Automatic Speech Segmentation

The next step is to align the phoneme sequence with the acoustic signal. Reliable automatic alignment/segmentation is also very critical for TTS design, i.e., manual segmentation is too labor-intensive to perform for hours of recording. We have used an automatic segmentation algorithm that adopts filter bank approach combined with wavelet convolution [9]. Preliminary evaluations indicate accuracy levels that, for most types of boundaries, are close to those of human segmentors. We also observe that even if the segmentor makes segmentation errors, most of the errors do not critically affect the performance of the proposed formant tracking algorithm.

Nominal (target) formant values [10] and voicing probability (1:voiced, 0:unvoiced and 0.3:mixed) are assigned to each temporal center of a phoneme segment. Formants and voicing probabilities for the frames between these center points are linearly interpolated.

Formant Tracking

The next step is to choose the best set of formant trajectories for N formants over K analysis frames. At each frame, k, there are L_k ways to map (assign) the candidate frequencies to formants. The L_k mappings can be identified as

$$L_k \;=\; \left(\begin{array}{c} n \\ N \end{array}\right) \;=\; \frac{n!}{(n-N)!N!}, \qquad (1)$$

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

The formants are chosen from the candidates based on minimal total cost, which is calculated from several cost functions: local cost, frequency change cost, and transition cost. The local cost λ_{kl}, of the l^{th} mapping at the k^{th} frame is based on the assigned bandwidths, B_{kln}, and the deviation from nominal formant frequencies for the phoneme, Fn_n,

$$\lambda_{kl} = \sum_{n=1}^{N} \left\{ \beta_n B_{kln} + \nu_n \mu_n \frac{|F_{kln} - Fn_n|}{Fn_n} \right\}, \qquad (2)$$

where β_n determines the cost of bandwidth broadening for the n^{th} formant, ν_n is the voicing probability and μ_n determines the cost of deviations from the nominal frequency of the n^{th} formant.

The frequency change cost, ξ_{kljn}, between the l^{th} mapping at frame k and the j^{th} mapping at frame $k-1$ for the n^{th} formant is defined as

$$\xi_{kljn} \;=\; \left\{ \frac{F_{kln} - F_{k-1jn}}{F_{kln} + F_{k-1jn}} \right\}^2. \qquad (3)$$

The quadratic cost function is to penalize any abrupt formant frequency change across analysis frames. Using Equation 3, a transition cost, δ_{klj}, can be defined as a weighted sum of the frequency change cost of individual formant:

$$\delta_{klj} \;=\; \psi_k \sum_{n=1}^{N} \alpha_n \, \xi_{kljn}, \qquad (4)$$

where α_n determines the relative cost of inter-frame frequency changes in the n^{th} 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 putting continuity constraints across those boundaries. The ψ_k can be any kind of similarity measures or inverse of distance measures such as inter-frame spectral distance measures in the LPC or cepstral domain. We use a simple stationarity measure based on the signal energy (rms), by which the weight of the continuity constraint can be reduced near the transient region. It is defined as the relative signal rms at the current frame:

$$\psi_k \;=\; \frac{rms_k}{\max_{i \epsilon K} \; rms_i}, \qquad (5)$$

with rms_k being the speech signal rms in the k^{th} analysis frame. 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 (Equation 2) is to prevent the 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:

$$C \;=\; \sum_{k=1}^{K} \min_{l \epsilon L_k} \; D_{kl}. \qquad (6)$$

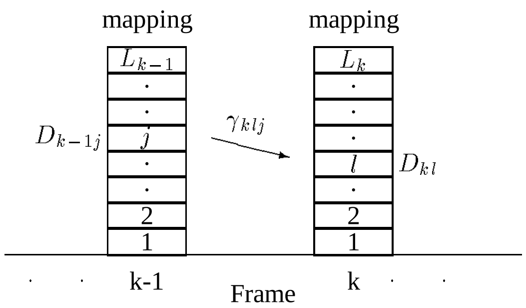

Figure 1: The mapping cost D_{kl} is the sum of local cost λ_{kl} and the minimum connection cost γ_{klj}. The γ_{klj} is calculated using the frequency change cost δ_{klj} and the mapping cost D_{k-1l} of the previous frame.

As shown in Figure 1 the mapping cost, D_{kl}, for the l^{th} mapping at the k^{th} frame is obtained from:

$$D_{kl} \;=\; \lambda_{kl} \;+\; \min_{j \epsilon L_{k-1}} \gamma_{klj}, \tag{7}$$

where λ_{kl} is given in Equation 2 and γ_{klj}, the connection cost from the j_{th} mapping at frame $k-1$ to the l^{th} mapping in frame k, is defined by the recursion:

$$\gamma_{klj} \;=\; \delta_{klj} \;+\; D_{k-1\,j}. \tag{8}$$

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 [7], while the value of μ_n is varied to find the optimal weight for the cost of deviation from the nominal formant frequencies.

RESULTS

The algorithm has been tested on 276 sentences spoken by a male speaker of American English. The speech corpus was originally created for the purpose of constructing an acoustic inventory for a concatenative TTS system. Each utterance has a carrier phrase and one or two target units (diphone or triphone) in the middle of the phrase. Formant tracking errors were visually inspected by overlaying the formant tracks on the corresponding spectrogram. Only the formant tracks near the vowel region of the target units were considered in calculating the tracking error rate. The evaluation was performed to determine the accuracy with which the best set of formant candidates is chosen. Thus, the absolute formant frequency accuracy was not of interest to us.

The performance of the new algorithm was compared with a formant tracker using only the continuity constraints. Formant tracking errors were labeled based on the following rules. If a tracker missed the first formant and, therefore, 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 higher 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.

Method	Errors (%)	F_1 errs	F_2 errs	F_3 errs	μ_n
CC	36 (13.04)	10	21	36	
P1	11 (3.99)	1	3	10	10
P2	10 (3.62)	1	9	10	7
P3	10 (3.62)	0	10	10	4

Table 1: Summary of formant tracking error for vowel-like sounds. Total errors are the number of utterances that have formant errors out of 276 test utterances (any formant error regardless of F_1, F_2 or F_3). The next three columns, F_1 through F_3 errors, show how each formant error was distributed over formant number. Since a formant error can happen at both F_1 and F_2, the first three formant errors do not add up to the Total Errors.

Table 1 lists the number of formant tracking errors. The first row, denoted as CC, shows the results for the formant tracker using the continuity constraints only. The next three rows P1, P2, and P3 are for the newly suggested algorithm with different weightings μ_n on the cost function (Equation 2). Smaller μ_n means less cost for deviation from the nominal formant values, resulting in relatively stronger continuity constraints. The best performance was obtained when μ is 7 or 4, though the difference is not very big.

As it would be expected, the new proposed algorithm gives much better results (less than 4% error rate) than the formant tracker CC (13.04% error rate). Notice that for the CC method a large portion of the errors are at F_1 (10/36=27.78%) and F_2 (21/36=58.33%), which is serious because these formants are more heavily weighted in the acoustic unit selection process than F_3. On the other hand, over 90% (10/11=90.9%, 9/10=90%, and 10/10=100% for three tests, respectively) of errors made by the new proposed algorithm occurred in the F_2 or F_3 track. The mismatch in the third formant is less penalized in the acoustic unit selection process.

Figure 2-4 shows an example of formant tracking results using both methods. The CC method (Figure 3) clearly missed the second formant track near the diphthong /ɔɪ/ segment (indicated by an arrow). It is probably because the continuity constraints forced the tracking algorithm to make the second formant in the /ɔɪ/ segment continuous to the second formant of the previous voiceless fricative /h/ near 2400 Hz. This is a typical example of failure, where the continuity constraints put too much emphasis

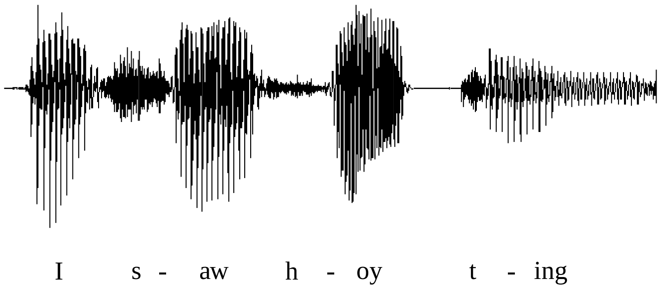

Figure 2: Speech waveform in “I saw hoyting guys”.

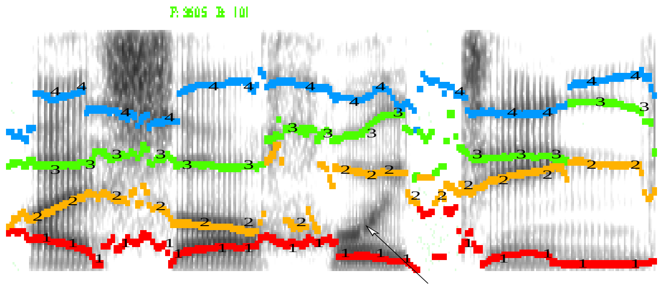

Figure 3: Spectrogram and formant tracks - CC.

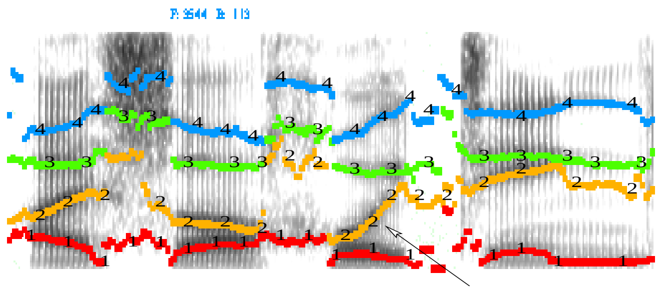

Figure 4: Spectrogram and formant tracks - proposed method.

on connecting the formant tracts of the vowel segment to the preceding fricative segment. Figure 4 shows the correct tracking results by using the proposed method, where the new algorithm found the second formant near the nominal formant values at about 1250 Hz of the /ɔɪ/.

In summary, although the current test data is spoken by only one male speaker, the above results indicate that once a nominal formant table for a given speaker is available, formant tracking performance can be much improved. For tests with a greater variety of speakers, separating nominal formant tables for different gender and age groups will be more effective.

DISCUSSION AND FUTURE WORK

We presented the implementation of new formant tracking algorithm using the knowledge of phonemic identity of the analysis frame. The new algorithm significantly reduced the error rate (3.62%) over the formant tracking algorithm using continuity constraints only(13.04%).

In general the new formant tracking algorithm is quite robust to small segmentation errors. However, errors tend to occur when there is severe coarticulation. For example, when a vowel /a/ is followed by a retroflex sound /r/ as in a diphone /a-r/, the formant tracks in the early part of /a/ often show the second formant around 1200 Hz, which is the second formant of /r/. Both methods often made errors in detecting the low second formant introduced by the following /r/ sound. This problem can be somewhat resolved by reducing the weighting factor μ in the Equation 2 such that the procedure becomes less sensitive to the phoneme boundary . A more systematic solution to this problem is to incorporate context dependent nominal formant values. This can be extended to allow alternate nominal formant values depending on the segmental context.

REFERENCES

1. R. Sproat, editor, (1998), *Multilingual text-to-speech synthesis: The Bell Labs Approach*, Kluwer Academic, Dordrecht; Boston; London.
2. R.W. Schafer and L.R. Rabiner, (1970), “System for automatic formant analysis of voiced speech,” *Journal of the Acoustical Society of America*, 57(2):634–648.
3. S. McCandless, (1974), “Automatic formant extraction using linear prediction,” *Journal of the Acoustical Society of America*, 54(1):339.
4. M. Hunt, (1985), “A robust formant-based speech spectrum comparison measure,” In *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing*, pages 1117–1120.
5. G.E. Kopec, (1986), “Formant tracking using hidden markov models and vector quantization,” *IEEE Transactions on Acoustics and Speech Signal Processing*, 34(4):709–729.
6. S. Seneff, (1986), “An auditory-based speech recognition strategy: Application to speaker-independent vowel recognition,” In *Proc. of Speech Recognition Workshop*.
7. D. Talkin, (1987), “Speech formant trajectory estimation using dynamic programming with modulated transition costs,” Technical Report 11222-870720-07TM, AT&T Bell Laboratories.
8. R.W. Hamming, (1962), *Numerical Methods for Scientists and Engineers*, McGraw-Hill.
9. J.P.H. van Santen, R. Sproat, (1999), “High-accuracy automatic segmentation,” *Proceeding of EuroSpeech99, Budapest Hungary*.
10. J.P. Olive, A. Greenwood, J. Coleman, (1993), *Acoustics of American English Speech - A Dynamic Approach*, Springer.