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). Formant 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 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

\[
L_k = \binom{n}{N} = \frac{n!}{(n-N)!N!} \tag{1}
\]

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 \(\lambda_{k1}\), of the \(l^{th}\) mapping at the \(k^{th}\) frame is based on the assigned bandwidths, \(B_{k1}\), and the deviation from nominal formant frequencies for the phoneme, \(F_{n_1}\),

\[
\lambda_{k1} = \sum_{n=1}^{N} \left( \beta_n B_{k1n} + \nu_n \eta_n \frac{|F_{k1n} - F_{n_1}|}{F_{n_1}} \right) \tag{2}
\]

where \(\beta_n\) determines the cost of bandwidth broadening for the \(n^{th}\) formant, \(\nu_n\) is the voicing probability and \(\eta_n\) determines the cost of deviations from the nominal frequency of the \(n^{th}\) formant.

The frequency change cost, \(\xi_{k1jn}\), 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

\[
\xi_{k1jn} = \left( \frac{F_{k1n} - F_{k-1jn}}{F_{k1n} + F_{k-1jn}} \right)^2 \tag{3}
\]

The quadratic cost function is to penalize any abrupt formant frequency change across analysis frames. Using Equation 3, a transition cost, \(\delta_{k1j}\), can be defined as a weighted sum of the frequency change cost of individual formant:

\[
\delta_{k1j} = \psi_k \sum_{n=1}^{N} \alpha_n \xi_{k1jn}, \tag{4}
\]

where \(\alpha_n\) determines the relative cost of inter-frame frequency changes in the \(n^{th}\) formant. The term, \(\psi_k\) 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 \(\psi_k\) 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:

\[
\psi_k = \frac{rms_k}{\max_i K rms_i}, \tag{5}
\]

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:

\[
C = \sum_{k=1}^{K} \min_{|l|\in L_k} D_{k1}. \tag{6}
\]
Figure 1: The mapping cost $D_{kl}$ is the sum of local cost $\lambda_{kl}$ and the minimum connection cost $\gamma_{ij}$. The $\gamma_{ij}$ is calculated using the frequency change cost $\delta_{kl}$ and the mapping cost $D_{k-l}$ of the previous frame.

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

$$ D_{kl} = \alpha_{kl} + \min_{j \in k-1} \gamma_{ij}, \quad (7) $$

where $\lambda_{kl}$ is given in Equation 2 and $\gamma_{ij}$, the connection cost from the $j^{th}$ mapping at frame $k-1$ to the $i^{th}$ mapping in frame $k$, is defined by the recursion:

$$ \gamma_{ij} = \delta_{ij} + D_{k-l}. \quad (8) $$

In the present implementation, the constants $\alpha_n$, $\beta_n$, and $\mu_n$ are independent of $n$. The values of $\alpha_n$ and $\beta_n$ are determined empirically [7], while the value of $\mu_n$ is varied to find the optimal weight for the cost of deviation from the nominal formant frequencies.

RESULTS

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>
<thead>
<tr>
<th>Method</th>
<th>Errors (%)</th>
<th>$F_1$ errs</th>
<th>$F_2$ errs</th>
<th>$F_3$ errs</th>
<th>$\mu_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>36 (13.04)</td>
<td>10</td>
<td>21</td>
<td>36</td>
<td>10</td>
</tr>
<tr>
<td>P1</td>
<td>11 (3.99)</td>
<td>1</td>
<td>3</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>P2</td>
<td>10 (3.62)</td>
<td>1</td>
<td>9</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>P3</td>
<td>10 (3.62)</td>
<td>0</td>
<td>10</td>
<td>10</td>
<td>4</td>
</tr>
</tbody>
</table>

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 $\mu_n$ on the cost function (Equation 2). Smaller $\mu_n$ means less cost for deviation from the nominal formant values, resulting in relatively stronger continuity constraints. The best performance was obtained when $\mu$ 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 $\text{iu}$ segment (indicated by an arrow). It is probably because the continuity constraints forced the tracking algorithm to make the second formant in the $\text{iu}$ segment continuous to the second formant of the previous voiceless fricative $\text{l}$ near 2400 Hz. This is a typical example of failure, where the continuity constraints put too much emphasis
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 $\mu$ 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