Estimating the Spectral Envelope of Voiced Speech Using Multi-frame Analysis

Citation for published version:

Published In:
Eurospeech 2003 - Interspeech 2003
We first find the mean phase-spectrum, $\varphi(f)$, of all the harmonics by taking a moving average of all the observed phases.

$$\varphi(f) = \frac{\phi(f)}{|\phi(f)|} \quad \phi(f) = \frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{N_k} \hat{v}_k G(f - f_k) e^{j(\omega(f) - 2\pi f_k)}$$

$\hat{v}_k$ : observed (wrapped) phases
$\tau_k$ : delay of frame $k$
$G$ : moving average window
$\nu_k$ : weight for harmonic $k$ of frame $k$

The phase $\varphi(f)$ is then adjusted using the delay $\tau_k$ so that the following equation is minimised.

$$\sum_{i=1}^{M} \sum_{k=1}^{N_k} \nu_k \left\{ \frac{\chi_k}{2\pi f_k} \right\}^2 = \sum_{i=1}^{M} \sum_{k=1}^{N_k} \nu_k \left\{ \frac{1}{2\pi f_k} \text{ARG} \left[ \frac{\varphi(f)_i}{\varphi(f)} \right] \right\}^2$$

We can obtain the correcting value $\Delta \tau_k$ for the delay $\tau_k$ as:

$$\Delta \tau_k = \sum_{i=1}^{M} \sum_{k=1}^{N_k} \nu_k \text{ARG} \left[ \frac{\varphi(f)_i}{\varphi(f)} \right] / \sum_{i=1}^{M} \nu_k$$

5. Results and discussion

- **MOCHA-TIMIT corpus**
  - number of sentences: 460
  - sampling rate: speech 16.0 KHz, DMSA 500 Hz

- 87208 voiced frames were extracted from the corpus using the following analysis.

  **Harmonic estimation**
  - method: weighted LSM (Stylianou, 2001)
  - analysis type: Hamming
  - window width: 210 ms
  - spacing: 8.0 ms

- **LBG clustering** (Linde et al., 1980) was applied to the articulatory data in order to identify frames with similar articulator settings. Cepstrum coefficients were then calculated by performing MFCA for all the frames in each cluster.

```
<table>
<thead>
<tr>
<th>Multi-frame analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of clusters</td>
</tr>
<tr>
<td>order of cepstrum</td>
</tr>
</tbody>
</table>

$w_k$ : $\exp \left\{ -\frac{\|f_k - \hat{f}_k\|^2}{2N_k \nu_k} \right\}$
$v_k$ : $\exp \left\{ -\frac{\|f_k - \hat{f}_k\|^2}{2N_k \nu_k} \right\}$
$G(x)$ : $\exp \left\{ -\frac{\|x\|^2}{2\sigma^2} \right\}$
```

5. Discussion

- MFCA discovers smooth spectral envelopes which best approximate all the harmonics of all the frames within each articulatory cluster.
- Some clusters have comparatively large variances of observed harmonic spectra, which is probably because:
  - each frame has a different noise level (S/N) that changes depending on speech powers of the frames;
  - we do not take into account the variance in the acoustic space during the clustering in the articulatory space;
  - the voice source is not a periodic impulse train as we assumed, and its spectral characteristic changes depending mainly on $f_0$ and power.

6. Future work

- In respect of the source problem, we have proposed an approach that can take into account the change of the voice-source characteristic using MFCA. (See Poster 3)
- The combination of MFCA and the articulatory clustering produces a codebook which relates articulation with acoustic feature of speech. With this codebook we are currently examining articulation-to-speech conversion, in which speech can be modified in articulatorily-meaningful ways.
Estimating the Spectral Envelope of Voiced Speech Using Multi-frame Analysis

Yoshinori SHIGA and Simon KING
Centre for Speech Technology Research
University of Edinburgh
yoshi@cstr.ed.ac.uk

1. Motivation

**Problem in spectral envelope estimation**
The spectrum of voiced speech only has energy at frequencies corresponding to integral multiples of \( F_0 \) and therefore it is impossible to identify the transfer characteristics between the harmonics.

**Conventional methods**
Spectral envelope estimation is interfered with by the harmonic structure in conventional methods.

2. The idea

**Using multiple speech-frames**
Using multiple speech signals generated with different \( F_0 \) through the same transfer function allows us to estimate more exact envelopes.

**How to collect the frames**
- Changing \( F_0 \) with fixed articulation
- High \( F_0 \)
- Medium \( F_0 \)
- Low \( F_0 \)

3. Multi-frame analysis

**Multi-frame cepstral analysis (MFCA)**
We employ the cepstrum as an expression of the spectral envelope.

**Estimating the envelope of amplitude spectrum**
To find a cepstrum which best approximates the amplitude of all the harmonics in multiple frames we use least squares estimation in the frequency domain, which is an extension of the method by Galas and Rodet (1990).

The sum of the squares of approximation error for all the harmonic amplitudes of all the frames is as follows:

\[
\sum_{k=1}^{M} \sum_{l=1}^{N_k} (|d_k| - |c_k|)^2 = \sum_{k=1}^{M} \sum_{l=1}^{N_k} c_k \left| u_k^H - d_k - \sum_{n=0}^{N_k} c_n \cos(2\pi f_n T) \right|^2
\]

- \( d_k \) : observed log-amplitude of harmonic \( l \) in frame \( k \)
- \( f_n \) : observed frequency of harmonic \( l \) in frame \( k \)
- \( c_k \) : cepstral coefficients
- \( N_k \) : number of harmonics in frame \( k \)
- \( M \) : number of frames
- \( T \) : sampling period
- \( u_k \) : weight for harmonic \( l \) in frame \( k \)
- \( d_k \) : offset that adjusts the total power (\( a_k \))

This can be solved by reducing it to a problem of weighted least squares.

- In a large speech corpus, there must exist multiple speech portions produced using the same vocal-tract shapes.
- Those portions usually have different \( F_0 \)s.
- Electro-magnetic articulograph (EMA) data indicates where identical vocal-tract shapes are in the corpus.
- Otherwise, phonetic information, such as phonemic context, could be used to identify those locations.