Comparison of Formant Enhancement Methods for HMM-Based Speech Synthesis

Tuomo Raitio[1], Antti Suni[2], Hannu Pulakka[1], Martti Vainio[2], and Paavo Alku[1]

[1]Aalto University, Department of Signal Processing and Acoustics, Helsinki, Finland
[2]University of Helsinki, Institute of Behavioural Sciences, Helsinki, Finland

Introduction

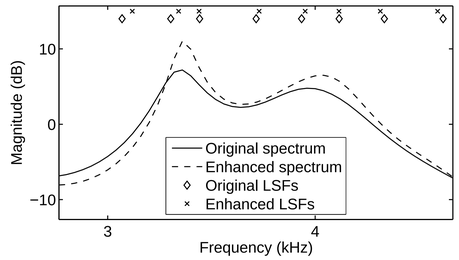
- Hidden Markov model (HMM) based speech synthesis has a tendency to over-smooth the spectral envelope of speech → Speech sounds muffled and unnatural
- One way to compensate for the over-smoothing is post-filtering → Enhance the dynamics between the formant peaks and the spectral valleys. This is also called as *formant enhancement*
- This study compares two formant enhancement methods: LSF-based method, and a new LPC-based method
- Formant enhancement before HMM-training is also studied

LSF-Based Formant Enhancement Method

- Line Spectral Frequency (LSF) based enhancement introduced by Ling *et al.*
- Based on modifying the LSF positions

→ Shift the LSFs that are close to each other even closer, which makes the spectral peaks sharper

- Formant positions are also slightly shifted
- Example of the procedure is shown in the left figure

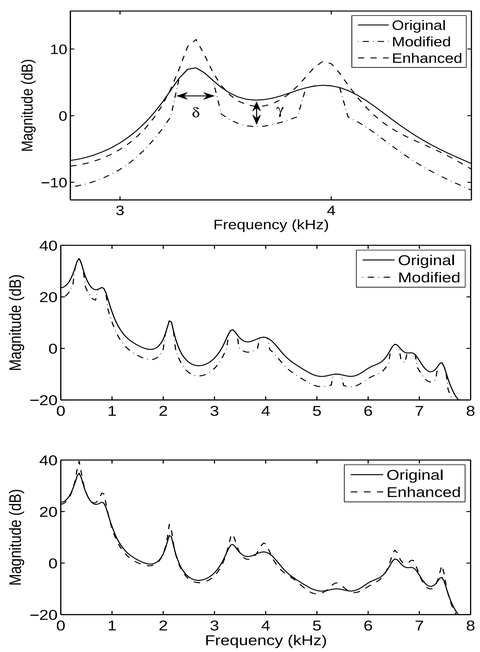


New LPC-Based Formant Enhancement Method

- New technique, referred to as the LPC-based formant enhancement method, is based on modifying the power spectrum of the LPC model and then re-evaluating new LPC. The algorithm:

1. Evaluate power spectrum from LPC (use FFT)
2. Modify the power spectrum by decreasing low-energy parts
   → Find formants and valleys from the smooth power spectr.
   → Multiply the low-energy regions by a small coefficient
   → Formants are left unmodified
3. Construct autocorr. from the new power spectr. (use IFFT)
4. Re-evaluate new LPC from autocorr. function (Yule-Walker)

- Since LPC analysis focuses on spectral peaks, the new LPC model will show sharper resonances
- The reduction in the low-energy parts is controlled by two parameters: The width δ of the unmodified area withing a spectral peak and coefficient γ ($0 \leq \gamma \leq 1$) that reduces the low-energy areas
- The parameters are shown in the upper figure
- Lower figure shows an example of the procedure



Formant Enhancement Prior to HMM Training

- Conventionally, the averaging effect is compensated for after the speech parameter generation
- This paper also studies the enhancement of formants prior to HMM training
- Pre-enhancement provides formant information that has higher dynamics
- More prominent formant information may yield more robust models and enhance synthesized speech

Experiments – Objective Evaluation

- Performance of the two formant enhancement methods (LSF and LPC-based) were studied objectively by measuring the bandwidth ratio ($R = B_{\text{enh}}/B_{\text{orig}}$) and formant shift ($\Delta F$) of the first two formants
- Database of eight Finnish vowels [*a*, æ, e, i, o, ø, u, y] spoken by ten Finnish speakers (5 males and 5 females)

- Similar bandwidth ratios
- LPC-based method has considerably lower formant shift
- Results differ from the ones in the paper (paper has faulty results)

- Bandwidth ratio (R)

Method	F1	F2
LSF-based-03	0.46	0.43
LSF-based-04	0.52	0.49
LSF-based-05	0.59	0.56
LPC-based-02	0.36	0.35
LPC-based-03	0.45	0.45
LPC-based-04	0.54	0.53

- Formant shift (ΔF)

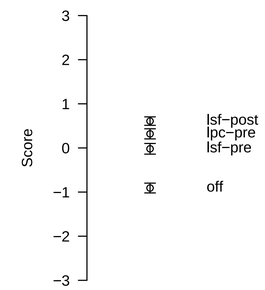
Method	F1 (%)	F2 (%)
LSF-based-03	3.40	1.94
LSF-based-04	2.85	1.72
LSF-based-05	2.32	1.48
LPC-based-02	0.84	0.36
LPC-based-03	0.78	0.38
LPC-based-04	0.71	0.37

Experiments – Subjective Evaluation

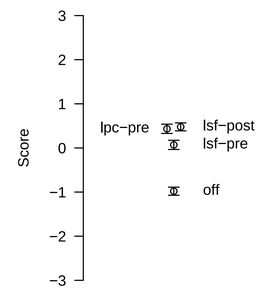
- Formant enhancement methods were evaluated subjectively by assessing the quality of synthetic speech of four different systems:
  1. No formant enhancement (off)
  2. LPC-based pre-enhancement (lpc-pre)
  3. LSF-based pre-enhancement (lsf-pre)
  4. Post-filtering with LSF-based enhancement (lsf-post)
- 1 hour of training speech material, CCR test, 11 test subjects, 2 test setups:
  - 1. Overall performance of synthetic speech
  - 2. Performance with normalized duration and F_0

  → test the quality of the formants

1. Overall performance



2. Normalized duration and F_0



- Overall, post-filtering was assessed better than pre-enhancement
- LPC-based method was assessed better than LSF-based method

Conclusions

- Pre-enhancement effectively alleviates the over-smoothing
- LPC-based method performed better in pre-enhancement compared to LSF-based method

tuomo.raitio@tkk.fi, antti.suni@helsinki.fi