Prosody Modification of Speech on Mobile Devices

BTP Report-I

Project Mentor: Dr. Kishore Prahallad

By,
Varshanjali Sayyaparaju (201030097)
Sravya Kanmanthareddy (201030155)

Introduction:

Speech prosody is a very common technique used in speech processing, for various applications such as voice conversion, speech synthesis, etc..

We are using speech prosody to make an application (on mobile devices), that employs prosody modification algorithms including duration, intonation and spectral modifications. This project deals with faster and efficient implementation of such algorithms.

By modifying duration, intonation and/or spectrum of speech, we are changing the voice characteristics, so that the output voice is different from the input voice. For example, in the application TomCat, our voice is given as input, and the output is the same speech in a cat's voice.

Our goal is to concentrate on **spectral changes** of speech to employ prosody modification and to develop a real-time efficient prosody modification application on android mobiles.

Theory:

Speech: Output of a time-varying vocal tract system excited by a time-varying excitation.

- Excitation source: The air rushing from the lungs past the vocal folds through windpipe.
- Vocal tract response: Shape of the vocal tract at different time instances.

Phone: The basic sound unit of speech, which consists of a consonant and a vowel. Ex: 'va', 'sa'.

Prosody: Is a characteristic of speech containing information on rhythm, stress and intonation. Prosody may reflect various features of the speaker or the utterance: the emotional state of the speaker; the form of the utterance (statement, question, or command); the presence of irony or sarcasm; emphasis, contrast, and focus; or other elements of language that may not be encoded by grammar or choice of vocabulary.

Prosody modification includes modification in:

- Duration: How long each phone should be. For example, the length of a phone changes based on emotional state of the speaker. It can be shorter when angry, compared to when normally said.
- Pitch contour: Fundamental frequency of speech, which varies from speaker to speaker. Ex: Female have a higher pitch, while a males is relatively lower.
- **Spectral changes:** Spectrum consists of all the characteristics of speech. One can obtain speech given its spectrum. Modifications in the spectrum reflects prosody modification of speech.

Current Scenario:

Popular algorithms used for Prosody modification are:

- TD-PSOLA
- Sinusoidal models
- Harmonics + noise models
- STRAIGHT

These algorithms are for pitch, and duration modification, we are interested in spectral modification. Therefore, we use:

- **LP Spectrum Modification using Pole-Zero Plot**
- Mel-Cepstral
- LP-PSOLA
- Frequency-domain PSOLA

Approaches:

RESAMPLING

Speech signals are generally processed in digital representation. For digitalization, we sample the speech signal at a sampling frequency of usually 8000Hz. In resampling, we change the sampling rate to create prosody modification.

Procedure:

- We recorded multiple vowels at a sampling frequency of 8000Hz
- We changes the sampling rate to 4000Hz for each vowel and observed changes
- Then, repeated the above step for 12000Hz and observed changes.

Observations:

- Resampling changes the voice characteristics
- But rate of sampling can not be fixed because distortion is noticed for some sounds,and not others. This is clearly noticed in the higer pitch period of "aaa" vs "eee".
- Time duration decreases by a significant amount as pitch increases

Conclusions:

- Resampling is the easiest method to implement prosody, but not the most efficient method.

Pole-Zero Shifting:

THEORY:

Pole-Zero Plots:

A pole–zero plot is a graphical representation of a rational transfer function in the complex plane which helps to convey certain properties of the system such as:

- Stability
- Causal system / anticausal system
- Region of convergence (ROC)
- Minimum phase / non minimum phase

A pole-zero plot shows the location in the complex plane of the poles and zeros of the transfer function of a dynamic system.

By convention, the poles of the system are indicated in the plot by an X while the zeroes are indicated by a circle or O.

A pole-zero plot can represent either a continuous-time (CT) or a discrete-time (DT) system. For a CT system, the plane in which the poles and zeros appear is the s plane of the Laplace transform. In this context, the parameter s represents the complex angular

frequency, which is the domain of the CT transfer function. For a DT system, the plane is the z plane, where z represents the domain of the Z-transform.

In speech, we consider the system to be an all-pole system, in which the poles are represented in the form r*exp(j*θ).

- r represents the bandwidth of the system
- θ gives information about the resonance frequency.

Modifying r and θ changes the entire system, hence creating prosody modification.

LP Analysis:

The redundancy in the speech signal is exploited in the LP analysis. The prediction of current sample as a linear combination of past *p* samples form the basis of linear prediction analysis where *p* is the order of prediction. The predicted sample $\hat{s}(n)$ can be represented as follows,

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k.s(n-k)$$

where a_ks are the linear prediction coefficients.

LP residual is the prediction error e(n) obtained as the difference between the predicted speech sample $\hat{s}(n)$ and the current sample s(n).

$$e(n) = s(n) - \hat{s}(n)$$

LP Spectrum:

The LP spectrum provides smooth vocal tract spectral characteristics. This can be computed from the Fourier representation of the LP coefficients.

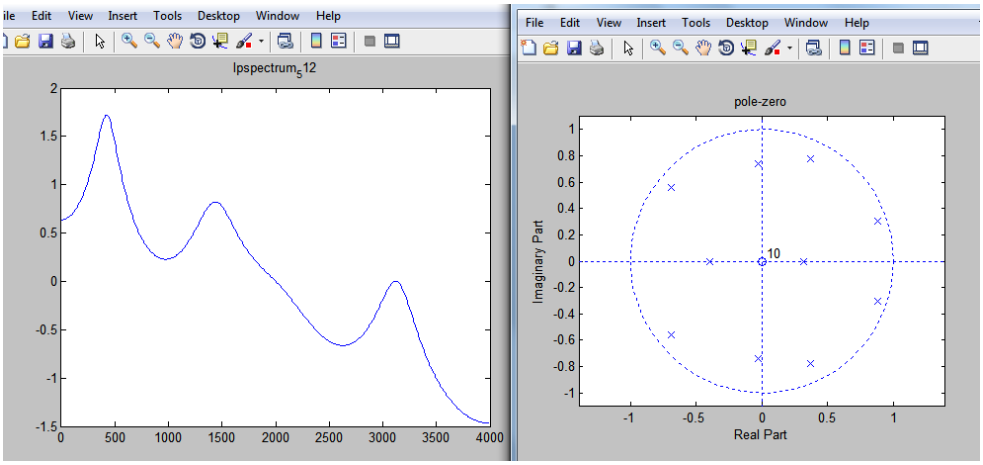
The basic shape of vocal tract can be characterized by the gross envelope of Linear Prediction (LP) spectrum. LP spectrum can be approximated by a set of resonant frequencies (formants) and their associated bandwidths. For each speaker, the shape of the vocal tract will be unique, and correspondingly the set of formants and their bandwidths. The resonances and their bandwidths are related to the angle and magnitude of the corresponding poles in the **z** -plane. The formant frequencies can be changed by shifting the poles of a system transfer function in the **z** -plane.

EXPERIMENT:

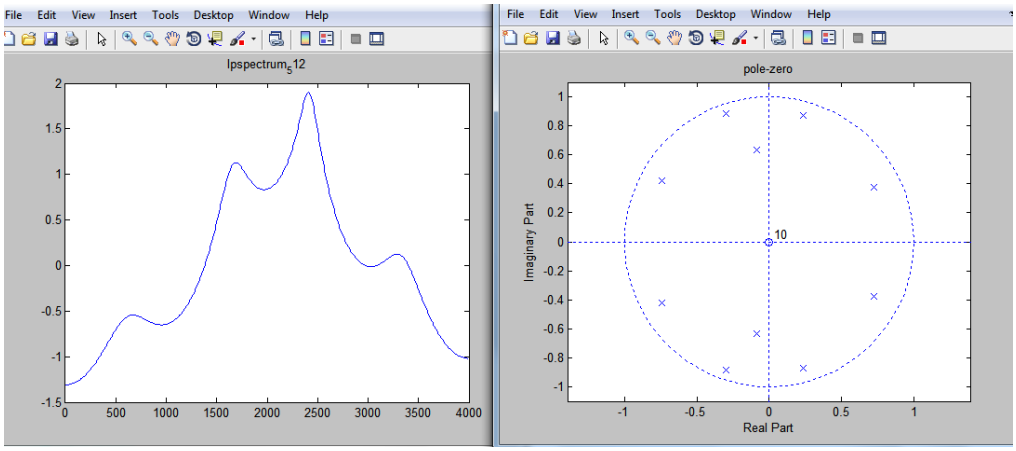
Analyzing TomCat:

We analyzed TomCat speech with that of human speech in the LP spectrum domain, and the corresponding Pole-Zero plots and observed changes in domain.

Input Voice:



TomCat Voice:



OBSERVATIONS:

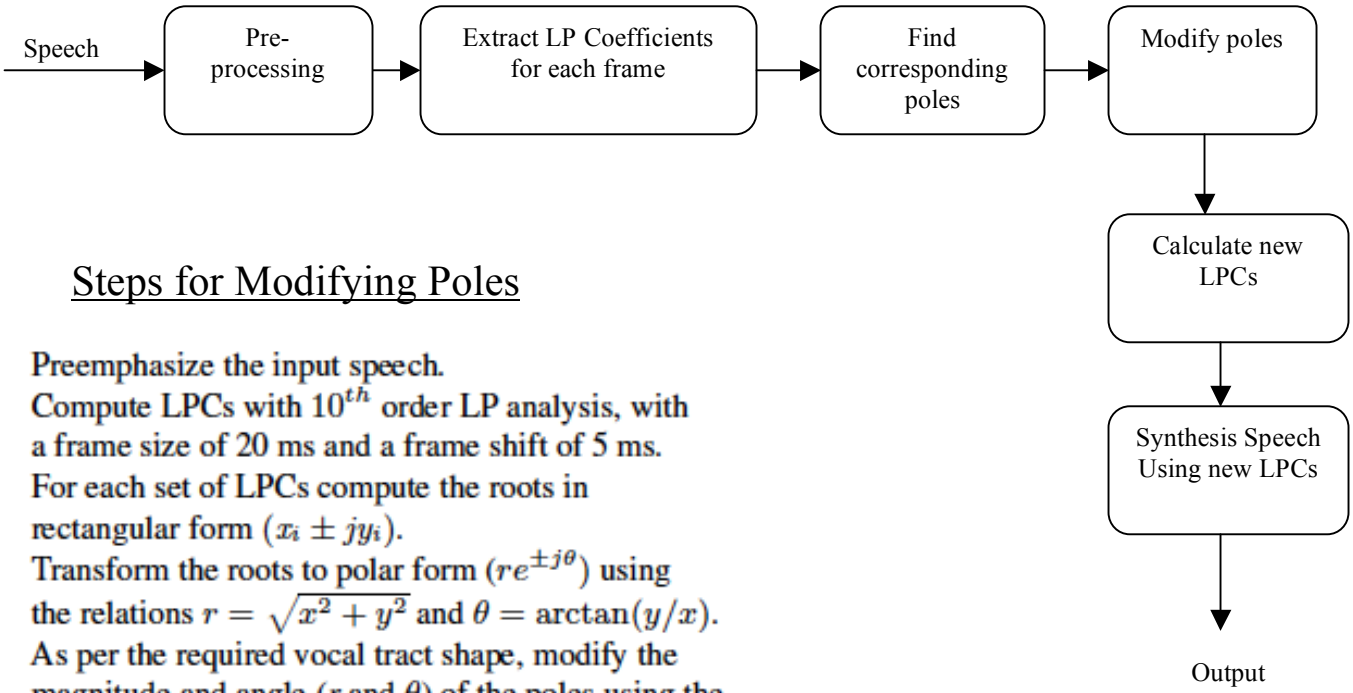
- Poles have been moved
- The spectrum has changed

CONCLUSION:

- Vocal tract characteristics can be changed by shifting poles of LP spectrum

LP Spectrum Modification:

Block Diagram:



Steps for Modifying Poles

1. Preemphasize the input speech.
2. Compute LPCs with 10^{th} order LP analysis, with a frame size of 20 ms and a frame shift of 5 ms.
3. For each set of LPCs compute the roots in rectangular form $(x_i \pm jy_i)$.
4. Transform the roots to polar form $(re^{\pm j\theta})$ using the relations $r = \sqrt{x^2 + y^2}$ and $\theta = \arctan(y/x)$.
5. As per the required vocal tract shape, modify the magnitude and angle (r and θ) of the poles using the relations $\theta_i' = \alpha_i\theta_i = \frac{f_i'}{f_i}\theta_i$ and $r = e^{-\pi\beta_i T}$. (where θ_i and θ_i' represents angular components of poles of source and target formant frequencies f_i and f_i', r = magnitude of the poles, β_i and T represents bandwidth of formants and sampling period.)
6. Transform the modified roots into complex conjugate form using the relation $(rcos\theta + jrsin\theta)$.
7. Compute the LPCs from the modified roots.

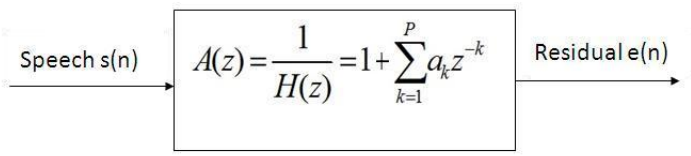
Changing the values of θ and r changes the vocal tract characteristics (formants position and bandwidth)

Speech Synthesis:

Inverse-Filtering

Method:

- Compute a_k's (LPCs)
- Pass speech through all-pole filter to obtain LP residual



- Modify poles
- Pass residual through all-pole filter using modified poles to obtain synthesized speech

RESULTS:

- As we decrease the value of θ keeping r constant, amplitude increases.
- As we increase the value of θ keeping r constant, amplitude decreases.
- Decreasing the value of r, dampens the signal, hence lower volume.
- Increasing r causes distortion, as well as increase in volume.

CONCLUSION:

- We observed that changing a single pole, causes less distortion in the signal. Therefore, we plan on continuing research to see:
  - How many poles we can change without distortion
  - To how much extent we can change the poles without distortion
- Meanwhile, for creating an android application, we need to employ prosody modification; therefore, we are looking into the domain of **FREQUENCY WARPING**, which we know will give required results.