Prosody Modification of Speech on Mobile Devices

BTP Report-II

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. 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.

We want to use the concept of ***Frequency Warping*** to implement spectral changes.

Frequency Warping:

Frequency Warping is a technique mapping one frequency axis to another using linear or non-linear functions. In our case, we use ***non-linear frequency warping***.

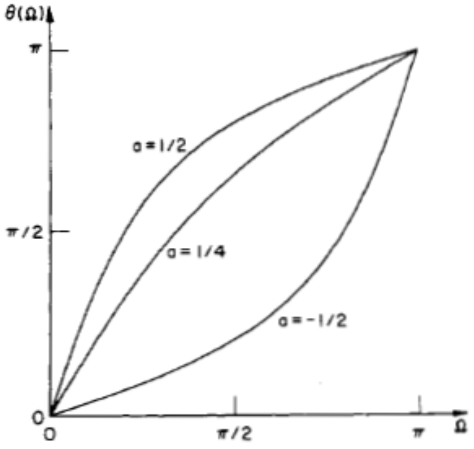


Fig. 1. Distortion of frequency for several values of *a*.

$$\omega = \theta(\Omega) = \tan^{-1}\left[\frac{(1 - a^2)\sin\Omega}{(1 + a^2)\cos\Omega - 2a}\right].$$

Ω(Original Frequency) is being mapped $\omega = \theta(\Omega)$ (Modified Frequency). For various values of parameter a, the mapping function varies. (When a = 0, the function becomes linear). Figure 1. is represented mathematically in the equation above.

The procedure consists of transforming the original sequence to a new sequence having the property that its DFT is equal to samples of the z transform of the original sequence at unequally spaced angles around the unit circle. Letting f(n) represent the original sequence, and g(k) the transformed sequence expressed in terms of a set of linearly independent sequences $\psi_k(n)$, so that :

$$f(n) = \sum_{k=-\infty}^{+\infty} g(k)\psi_k(n).$$

$$G(e^{j\omega}) = \sum_{k=-\infty}^{+\infty} g(k)e^{-j\omega k}$$

$$F(e^{j\Omega}) = \sum_{n=-\infty}^{+\infty} f(n)e^{-j\Omega n}$$

G and F are the Fourier Transform of the original and warped sequences respectively in their frequency domains.
And G and F are related in such a way that the frequency response G(ω) is corresponding to the value at F(Ω).

$$\omega = \theta(\Omega)$$

$$G(e^{j\theta(\Omega)}) = F(e^{j\Omega}).$$

Using the above equations, and including the conditions:

- The filter needs to be rational

We get:

$$\Psi_k(z) = \left(\frac{z^{-1} - a}{1 - az^{-1}}\right)^k.$$

This function represents a chain of filter order all-pass filters. For an all-pass filter, we know that the magnitude is unity, so only phase is being modified, by the equation:

$$\omega = \theta(\Omega) = \tan^{-1}\left[\frac{(1 - a^2)\sin\Omega}{(1 + a^2)\cos\Omega - 2a}\right].$$

Therefore, if given a sequence f(n), to get the sequence g(k), the filter that must be used is the inverse filter of the above filter, which is of the form:

$$H_k(z) = \begin{cases} \dfrac{(1 - a^2)z^{-1}}{(1 - az^{-1})^2}\left(\dfrac{z^{-1} - a}{1 - az^{-1}}\right)^{k-1}, & k > 0 \\ \dfrac{1}{1 - az^{-1}}, & k = 0. \end{cases}$$

This can be implemented as shown below:

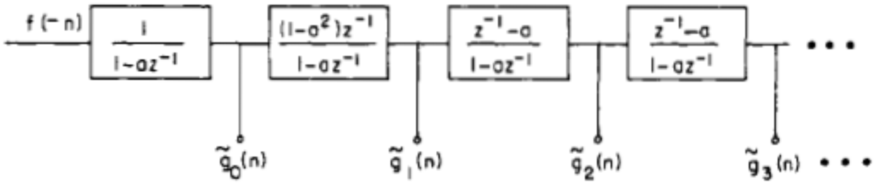


Fig. 2. All-pass network used to implement a distortion of the frequency axis.

Where the output sequence is taken after each individual filter, after processing all n input samples, i.e. g(k) = $g^{\sim}{}_{k}$(n).

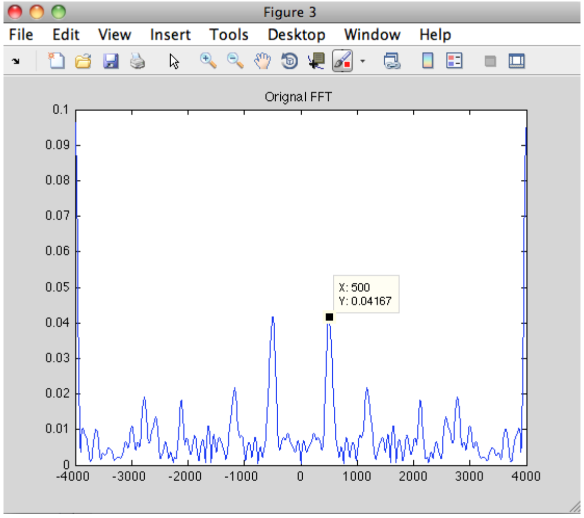
Implementation of Filter:

- Used Matlab
- Filter was implemented in time-domain taken in the form of linear-constant co-efficient difference equations.

Results:

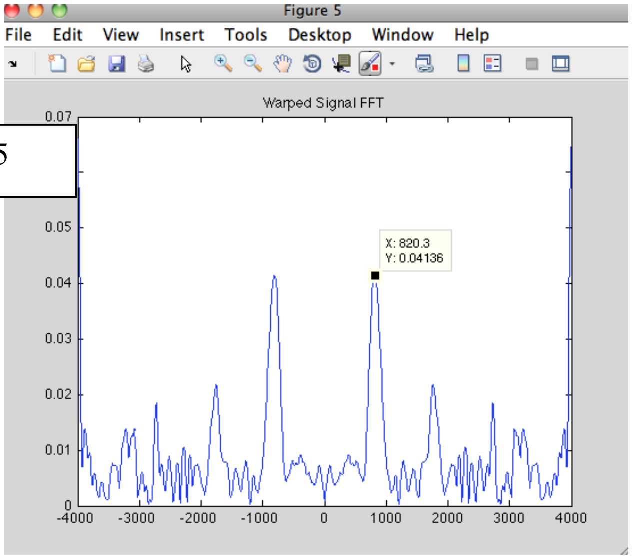
- When speech is passed through this filter, the required prosody modification has been achieved.

Spectrum of Original Speech

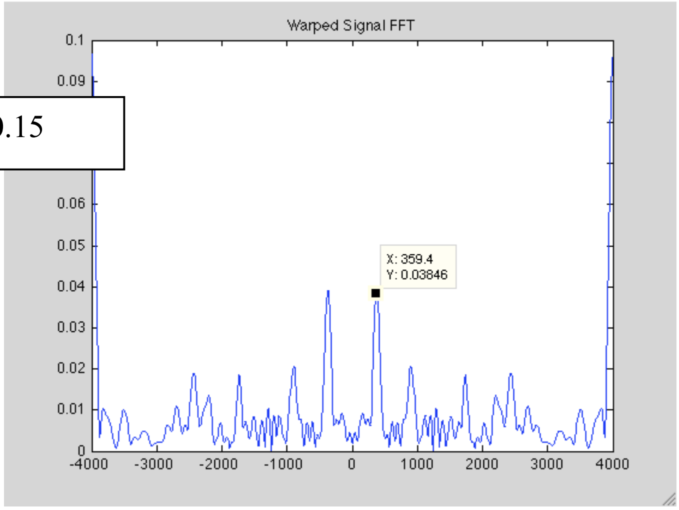


Spectrum of Warped Speech:

a = 0.25



a = -0.15



Variations in Experiments:

So far, we warped the speech signal directly. Speech system is composed of a vocal tract system represented as an impulse response of a filter, and an excitation source, which represented as residual. We found that modification of individual parameters produces prosody modification:

Conclusion:

1) Speech Warping:

Range of modification values with comprehensible female speech: [-0.15, 0.30]

Range of modification values with comprehensible male speech: [- 0.1,0.4]

Advantages: Efficient run-time

Disadvantages: Small range

2) Residual Warping

Range_female: [-0.9,0.9]

Range_male: [-0.9,0.9]

Advantages: Wider range

Disadvantages: Fluctuation in quality of speech within range.

3) Residual and Impulse Warping

Range_female: [-0.25,0.3]

Range_male: [-0.1,0.25]

Advantages: ---

Disadvantages: Inefficient run-time and small range

4) Impulse Warping

Range_female: [-0.3,0.25]

Range_male: [-0.1,0.3]

Advantages: ---

Disadvantages: ---

Observation:

-> 1) and 2) method are preferable.

-> A modifying factor can be fixed and an application can be made.