Male-female voice conversion

Stefano Cecconello, ceck1991@gmail.com, s1595539

January 12, 2015

This relation speaks about a male to female voice conversion program. The program is made of a training part and a conversion parts, that are explained in the Design paragraph. In the implementation I will introduce some parameters to adjust the final output. In the experiment section is exposed how to tune these parameters by empirical experiments using test on the output quality. In the conclusion I will demonstrate how the conversion quality could be considered satisfying with the usage of this method.

1 Introduction

In this document we will be discussing the issue of male to female voice conversion. The aim is to convert a given male sentence in a new sentence that seems like pronounced by a female.
The state of art in this field works not only on frequencies modify. Actual techniques focus also on manipulation of vocal apparatus. The aim is to synthesize a female voice based on the physical characteristics of the speaker.
My approach consists on a simple method based only on frequencies modification. The technique consists in two parts: the first is the learning part. Here the program tries to understand a function for frequencies modify. The second part consists on the conversion. The frequencies of a given male voice has been converted using the previous understand function.
The results of this approach give low sound quality voice conversion results, this is probably due to the simple method use for frequencies change. Anyhow the final output of the first part seems like we expect, also the final voice certainly moves in the direction of a female sound.

2 Design

The project is divided in two parts. The idea behind these is a conversion driven by a learning.
In the first part the program focuses on learning a conversion function starting from two given files. These are the same sentence, said by a man and by a

woman. The idea is to find the most significant frequencies for every window. Then a map has been made from male to female frequencies. By interpolation is generated a polynomial function from the previous points.
In the second part starting from the learned function the frequencies of a new given file are changed. After this conversion in the final output we obtain a new sentence that seems like said by a woman.

3 Implementation

In the training two files are required. They are divided in windows of the same dimension. The input files are supposed to be distort in time and in alignment at most in order of millisecond in every position.
A conversion from the time domain to the frequencies domain is then applied to every windows. Then for every frame of the male sentence the most important frequencies are extracted. This is done checking the power of every frequency. The strongest N frequencies of the window are took in account. Then these N frequencies are inserted in a table and one by one they are associated with the first N frequencies of the corresponding window in the female sentence. So in the table the most important frequency of the male sentence is associated with the most important frequency in the female sentence, the second with the second and so on. When this process has been done for every window then a regression algorithm is applied. From all the pairs is calculated a polynomial. It is important to underline the chose of a polynomial with a low degree. This choice was done because a high degree polynomial could perform worst than a low degree polynomial. This is because it is simple to fall in an overfitting problem using high degree polynomial. After this process we have a polynomial. As I said at the beginning of this chapter the sentences in the two audio files could be not perfectly aligned or could have different time length. This means that some pairs of equal letters, in the two audio files, could be or not be in the same windows. To remove this problem all the training process is repeated more than once, and every time a different window size is choose. With this strategy two letters that are not in the same window, the first time, can be in the same window with different window size. This avoid in part the alignment problem and the problem of different time length. At the end the polynomial is defined by doing the mean of all the found values of the polynomial for different window size.
The second part apply the polynomial at the frequencies of the target audio file. It is also apply a filter for don't introduce frequencies that are not in the range 20 Hz-20000 Hz.

4 Experiments

The experiments conducted are initiated by generating pairs of files aligned and with time length with difference at the order of milliseconds, how requested in the previous chapter. This was done using the program audacity.
In every experiment the modified parameters are: the number of different window sizes used, the windows size and the number of frequencies taken into account in every window. The parameters are evaluated by listening the final audio file and taking into account both audio quality and quality of the conversion. The training have been run on spoken sentences. The optimal values found for the parameters are:

- Number of different sizes of window used: 5. The conversion quality is worse using higher values. For this reason the number is small;

- Window size: from 24 to 40 ms. Bigger windows sizes have been tried but the quality of the final audio file go down very fast with higher size;

- Frequencies for window taken into account: 8. This number is completely empirical. Very low values for this number work very bad, but increase this value more then 8 gives lower conversion quality.

In general the polynomial, from the training part, seems good. This because for every input frequency an higher output frequency is returned. This seems correct because normally the frequencies of a female voice are higher of the frequencies of a male voice.
For conversion quality the results are not really good, but the final result seems like a sort of female voice. Also the quality, of final audio file, could be accept also if in the result is introduced some noise.

5 Discussion and Conclusion

From the result of the different experiments is possible to see that the technique works. The low quality is possible due to the to much simple techniques used for the conversion part. This conversion introduces in fact noise, this is due in part to the presence of noise in the original file. This noise is considered in the same way of the voice part of the audio file, so also these frequencies are modified during the conversion.
For the quality of the conversion the result could be acceptable. The voice changes in the direction of a sentence pronounced by a female speaker. A better conversion could be performed using more pairs for the training part and trying to understand better parameters.

In conclusion the program has reached final quality that can be considered satisfying considering the simplicity of the used techniques. New work can be done using more accurate methods for the frequency change and using a bigger set of training file. The possibility of increasing the degree of the used polynomial could also be considered however this involves to taking into account the possibility of an overfitting problem.