euronoise

On the use of scale transform in modeling of the spectral envelope of vowels

M. Jamaati, M. Lankarany and H. Marvi

Technical University of Shahrood, 12345 Shahrood, Iran
mahdi.jamaati@gmail.com

Vowel recognition has been commonly used in speech recognition system for large vocabulary continuous speech and isolated word recognition. This paper proposed a new feature extraction approach for vowel recognition using mellin transform and Mel Frequency Cepstral Coefficient (MFCC). The key property of mellin transform is the scale invariance which makes the features insensitive to different vocal tract lengths. The proposed idea combines the influence of the scale invariant property of mellin transform with the property of MFCC. On of the advantages of proposed algorithm in compare of previous algorithm is that there is no control parameter or tuning parameter in this method. Experimental results which have been done on a vowel database show a promising result.

1 Introduction

Speech recognition involves a number of stages. The first stage is to convert the analog waveforms of our speech into a digital form which computer can work with them. After digitization the data is simplified and analyzed into segments called speech representations. Speech recognition then involves matching the speech representation against the program's templates and then using both acoustic and language models to determine the most likely word that was spoken.

A fundamental question in the study of human auditory mechanisms is the following. Suppose that a man, woman, and child are asked to read aloud a sentence. How does the human auditory system process the information and identify that the same sentence has been spoken by all three? To answer this question, we first must tackle the simpler one of how the auditory system adjusts for the difference in body sizes and vocal tracts of the humans who are speaking. Irino and Patterson demonstrated that essential features in speech signals for the same vowel can be mapped consistently by the Mellin transform to the (nearly same) location in the time-scale plane regardless of the speaker's body size.

It is well known that the speech signal carries information about vocal tract length (VTL): for example, the formant frequencies of vowels decrease as the VTL increases. VTL normalization (VTLN) is now a commonly used normalisation technique in speech recognition. VTL has a substantial effect on the observed spectrum. for example, a typical female speaker exhibits formant frequencies around 20% higher than those of a male speaker. [2]

Vowel recognition has been commonly used in speech recognition system for large vocabulary continuous speech and isolated word recognition.

Vowels are produced by exciting a fixed vocal tract with quasi – periodic pulses of air caused by vibration of the vocal cords. The way in which the cross – sectional area varies along the vocal tract determines the resonant frequency of the tract (Formants) and thus the sound that is produced. Each vowel sound can be characterized by the vocal tract configuration that is used in its production. It is obvious that this is a rather impressive characterization because of the inherent difference between the vocal tracts of speakers. An alternative representation is in terms of the resonance frequencies of the vocal tract, but in this way a great deal of variation clearly exists in the vowels formant [3]. By using this new feature extraction approach, we can easily recognize vowel sounds independent of the speaker type, such as size, age and sex.

This paper is organized as follows. In section 2 and 3 definition of Mellin and scale transform are described, respectively. Fast mellin transform represented in section 4. A MFCC algorithm is presented in section5. The suggested method is introduced in section 6. Experiments results are shown in section 7, followed by a conclusion in section 8.

2 Mellin Transform

The Mellin transform is the most popular transform in the analysis of algorithms. It is closely related to the two-sided Laplace and Fourier transforms except that it has a polynomial kernel. The Mellin transform can represent a signal in terms of scale. The scale can be interpreted, similarly to frequency, as a physical attribute of signals. The proposed fast (subquadratic) implementation allows this transform to be used in practical applications. [1]

Mellin transforms are important in vision and image processing. In particular, a so-called Fourier-Mellin transform can be used for pattern recognition for its invariance to shift, scale, and rotation. The mellin transform is defined as follow:

$$M_f(p) = \int_0^\infty f(t)\, t^{p-1} dt \tag{1}$$

in the complex variable $p = -jc + \beta$, with $\beta \in \Re$, fixed parameter and $c \in \Re$ independent variable. We call this family of transforms the $\beta -$Mellin transform.

The existence of the Mellin transform (1) depends upon convergence of the transform integral.

$$\int_0^\infty \left| f(t) \right| t^{p-1} dt < \infty \tag{2}$$

This is a general sufficient condition for the existence of the transform. Further considerations can be made using the fact that $p = -jc + \beta$, and different or simpler forms of equation (2) can be derived. [4]

3 Scale Transform

A sound is a very complex phenomenon that can be represented in terms of air pressure over time. The operator method with the scale operator gives us the transform pair for the scale domain. The following transform integral is obtained:

$$D_f(c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t)\, e^{(-jc-1/2)\ln t} dt \tag{3}$$

where *c* is the scale variable. The scale inverse transform is given by

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} D_f(c) e^{(jc-1/2)\ln t} dc \qquad (4)$$

The transform (3) can be easily rewritten (when $p = -jc + 1/2$) and reconsidering the energy normalization parameter $\frac{1}{\sqrt{2\pi}}$) in the following form:

$$M_f(p) = \int_0^{\infty} f(t) t^{p-1} dt \qquad (5)$$

that is the Mellin transform. The scale transform is a restriction of the Mellin transform on the vertical line $p = -jc + 1/2$, with $c \in \Re$.. This transform is said to be scale – invariant (the Fourier transform is shift invariant), thus meaning that the signal differing just by scale transformation (compression or expansion with energy preservation) have the same transform magnitude distribution.

As already told, the key property of the scale transform is its scale invariance. This means that if f is a function and *g* is a scaled version of *f*, the transform magnitude of both functions is the same. A scale modification is a compression or expansion of the time axis of the original function that preserves signal energy. Thus, a function *g(t)* can be obtained with a scale modification from a function *f(t),* if $g(t) = \sqrt{\alpha} f(\alpha t)$ with $\alpha \in \Re$ and α > 0. Given a scale modification with parameter α, the scale transforms of the original and scaled signals are related by

$$D_g(c) = \alpha^{jc} D_f(c) \qquad (6)$$

This property derives from a similar property of the Mellin transform. In fact, if $h(t) = f(\alpha t)$, then

$$M_h(p) = \alpha^{-p} M_f(p) \qquad (7)$$

In both (6) and (7), scaling is reflected by a multiplicative factor for the transforms, and for (6) such factor reduces to a pure phase difference. So, the scale transform magnitude of the original signal of the scaled signal is the same. [5]

$$\left|D_g(c)\right| = \left|D_f(c)\right| \qquad (8)$$

The scale transform allows us to represent a signal in terms of scale components. This different prospective can help us to analyze and extrapolate different information that cannot be seen with the classical analysis. In general, in theory of wavelets, the scale with temporal information is a central concept. This is a desirable property if we want a joint representation, but if we need to analyze only the scale, the only solution is the use of the scale transform.

4 Fast Mellin Transform

Computing a discrete Mellin transform is relatively straightforward. For example, we can do an approximation of the transform integral using the Riemann sum. Unfortunately, doing this would give us algorithms exhibiting quadratic complexity, thus meaning that they are not usable in most practical applications. The basic idea of the fast Mellin transform (FMT) algorithm comes from (9), in particular when β = 1⁄2 (scale transform). While presented in prior works this idea is here used to build a practical and efficient computer program (in particular a Matlab toolbox). The algorithm approximates

$$M_f(c) = \int_{-\infty}^{\infty} f(e^t) e^{\beta t} e^{-jct} dt \qquad (9)$$

In the other words

$$D[f(t)] = F[e^{t\beta} . f(e^t)] \qquad (10)$$

Where F[.] and D[.] refer to the Fourier and scale transform, respectively. The equation suggests that we can calculate the scale transform using the Fourier transform after an appropriate warping on the signal. The block diagram is shown in figure 1.

The first step can be seen as an exponentially sampling of the continuous time signal. Since we have only a uniformly samples signal, we used a natural cubic interpolator, and have the linear complexity associated with resolution of a tridiagonal matrix. Using this interpolator we can resample the original function obtaining an exponentially sampled version. At the second stage multiplies it by an exponential, and finally performs a Fast Fourier Transform.

The algorithm has an asymptotic complexity that depends only on the FFT, as this is most complex part of the entire process. The asymptotic complexity of the entire process is $O(n \ln^2 n)$.The accuracy of the fast mellin transform in providing an approximation to the continuous- time mellin transform is good. The error due to the interpolation can be reduced using more samples (over sampling) in the warping process.

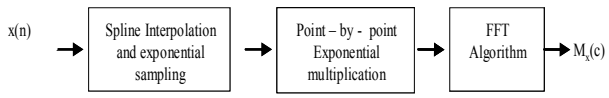


Figure 1 The fast mellin transform block.

5 Mel Frequency Cepstral Coefficient (MFCC)

Mel frequency Cepstral coefficient is the most popular feature extraction for speech recognition. MFCC are based on the known variation of human ear's critical bandwidth with frequency. The stages to calculated MFCC are as follow:

At the first step, divide signal into frame. Then for each frame, obtain the amplitude spectrum. And the next stage takes the logarithm and converts to Mel spectrum. And finally, the discrete cosine transform (DCT) is taken.

The block diagram of MFCC algorithm are shown in figure2,

6 Suggested Method

The block diagram of the suggested method is shown in figure3.

In the first stage of suggested method, the speech signals are broken into frame with 25.6 ms and update every 10 ms using hamming window. After that a new kind of cepstrum which apply the power function instead of log is used to obtain the spectral envelope of windowed signal. Figure 4.a show the vowel ‘’uu’’ pronounced by two different peoples. As it is show, the time domain waveforms of these vowels are different from each others. Figure 4.b show the improved cepstrum method of these two waveforms. As it is shown both these two curves are almost the same. Furthermore figure 5 show the time domain and the improved cepstrum of two different vowels. As it is clear, there is no similarity between the cepstrum of these two vowels.

In the second stage, the mellin transform of the spectral envelope are taken. The fast mellin transform of figure 4.a and figure 5.a are shown in figure 6.

In the third stage, Mel Frequency Cepstral Coefficient (MFCC), which described in section 5, are applied to feature extraction of the output of second stage. Figure 7 show the feature extracted from the third stage of proposed method which we are uses them in classification stage.

Classification of the vowels has been done in the last step. The purpose of this step is to determine to which class, a given input vowel sample belongs. This is based on a set of features extracted from third step, which make up the feature vector. The classifier uses these features to assign an input vowel to the correct class. A template matching involves a comparison of an average of features, computed on the test pattern, to a collection of stored average for each of the classes in training which is know as pattern. In our experiment a simple template matching, where the whole pattern is compared with a references pattern by measuring the Euclidean distance between features means are used.

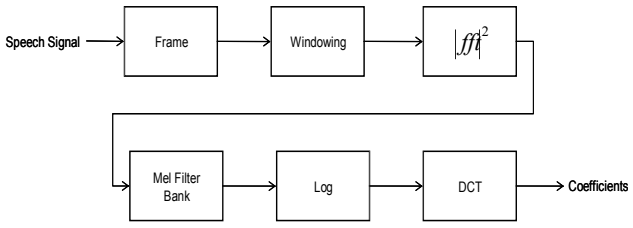


Figure 2 Block diagram of MFCC algorithm

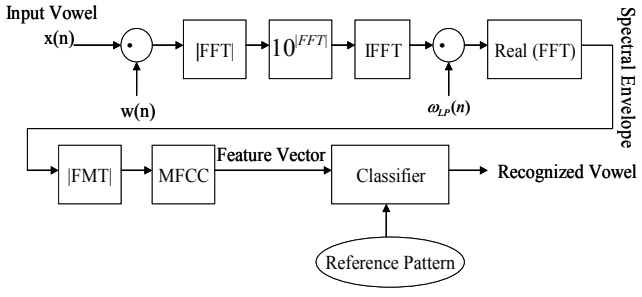
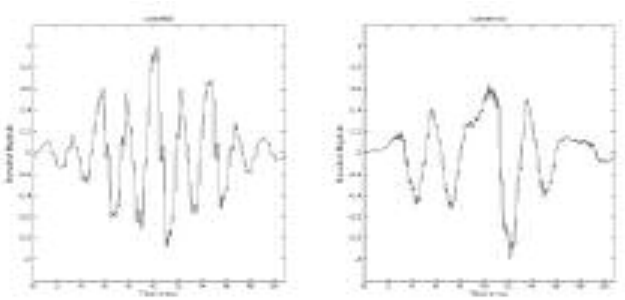
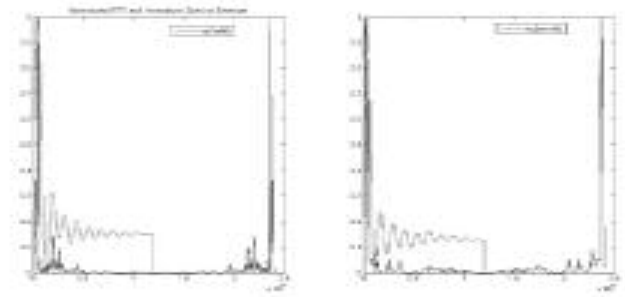


Figure 3 Block diagram of the suggested method

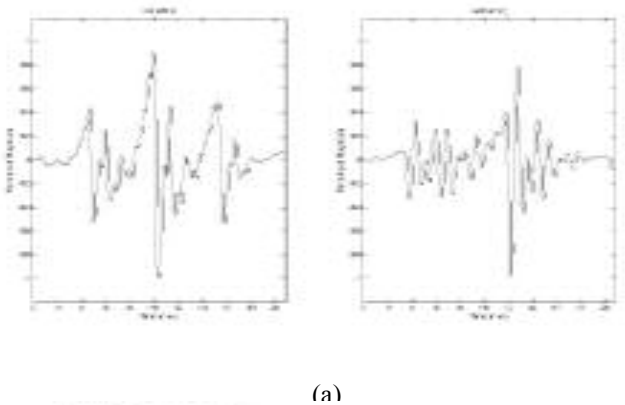


(a)

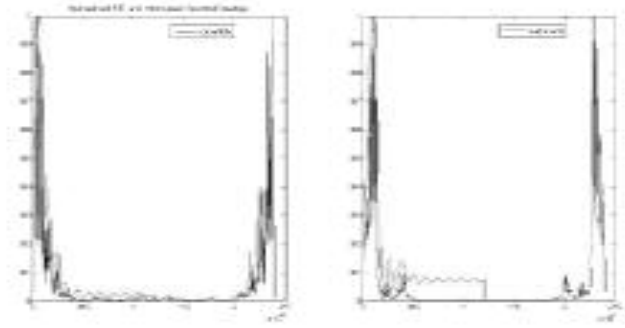


(b)

Figure4 (a) Vowel ‘’uu’’ pronounced by two different peoples. (b) The improved cepstrum method of these two waveforms.



(a)



(b)

Figure 5 (a) Time domain representation for two different vowels. (b) Improved cepstrum of these two vowels

7 Experimental results

Some experiments have been carrying out to evaluate the performance of suggested method in speech recognition and application. The database is a database of vowel

which are taken form []. This database is consist of 300 speakers including male, female, and chid. Each speaker pronounced 10 different vowels four times. Energy normalization has been done before any processing.

The proposed algorithm has been applied to each vowel to produce a feature vector as has been described in section 6.

By using of this algorithm, we can represent each vowel with only 13 parameters. For comparing of this parameter with reference parameters, the Euclidean distance has been used.

Table 1 show the accuracy rate of vowel recognition. As it is shown in this table, we can achieved an accuracy rate of %100 for three vowels. The accuracy rate for vowel of "aa" is %80 which is the lowest case.

Vowel	example	Accuracy rate
ii	beet	%100
uu	foot	%90
oo	bought	%100
aa	hot	%80
ee	bet	%100

Table 1 The accuracy rate of vowel recognition with suggested method

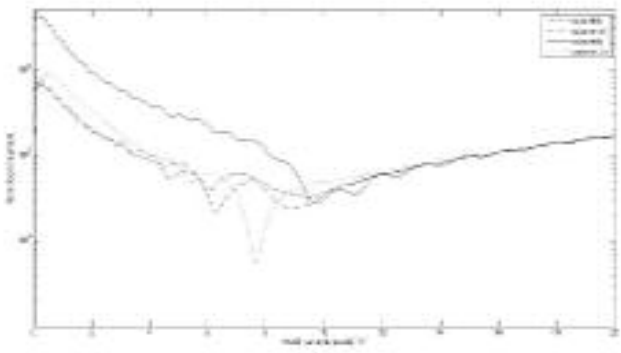


Figure 6 The fast mellin transform of figure4.a and figure 5.a

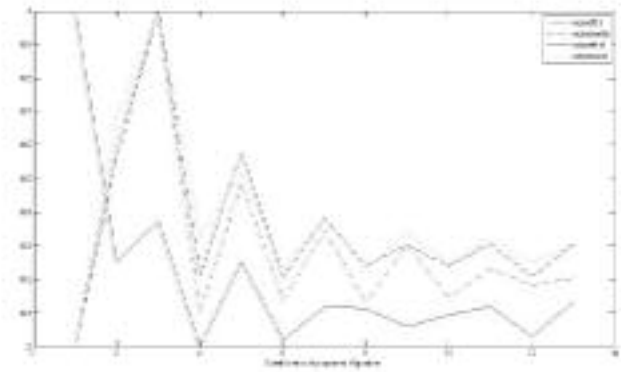


Figure 7 Feature extracted from the third stage of proposed method

8 Conclusions

An effective technique of feature extraction for vowel recognition is presented. The proposed idea is the influence of the key property of scale invariance of Mellin transform combined with the property of MFCC. In the experiments the performance of proposed feature has been evaluated. The advantage of proposed method in compare of previous algorithms is that there is no control parameter or tuning parameter in this method. Experiments which have been done on a vowel database show a promising result.

References

[1] Antonio De Sena and Davide Rocchesso, 2007. "A fast mellin and scale transform". EURASIP Journal on Applied Signal Processing. Volume 2007 , Issue 1 (Janurary 2007) Page: 75

[2] David R. R. smith and Roy D. Patterrson, 2005. "The interaction of glottal – pulse rate and vocal tract length in judgments of speaker size, sex, and age".

[3] L.R. Rabiner and R.W.Schafer . "Digital Processing of Speech Signal" .

[4] Antonio De Sena, Ph.D Proposal: "The Scale Domain".

[5] A. De Sena and D.Rocchesso, "A Fast Mellin Transform with Application in DaFx," in proc, of the int. conference on Digital Audio Effects (DAFx'04)