THE MEL-FREQUENCY CEPSTRAL COEFFICIENTS IN THE CONTEXT OF SINGER IDENTIFICATION

Annamaria Mesaros[1,2]
[1]Technical University of Cluj Napoca
Communications Department
Cluj Napoca, ROMANIA
annamaria.mesaros@com.utcluj.ro

Jaakko Astola[2]
[2]Institute of Signal Processing
Tampere University of Technology
Tampere, FINLAND
jaakko.astola@tut.fi

ABSTRACT

The singing voice is the oldest and most complex musical instrument. A familiar singer's voice is easily recognizable for humans, even when hearing a song for the first time. On the other hand, for automatic identification this is a difficult task among sound source identification applications. The signal processing techniques aim to extract features that are related to identity characteristics. The research presented in this paper considers 32 Mel-Frequency Cepstral Coefficients in two subsets: the low order MFCCs characterizing the vocal tract resonances and the high order MFCCs related to the glottal wave shape. We explore possibilities to identify and discriminate singers using the two sets. Based on the results we can affirm that both subsets have their contribution in defining the identity of the voice, but the high order subset is more robust to changes in singing style.

Keywords: sound source identification, singing voice, MFCC

1 INTRODUCTION

Considering the wide area of signals from our everyday life, problems concerning the singing voice characterization arise naturally after the interest on speaker and instrument recognition. Because of its particularities in production and control, the singing voice falls between the speech and musical instruments sounds, having common characteristics with each of these, but being also very different from both of them. Singing is composed mostly of sustained vowels with almost perfectly harmonic spectrum, resembling with the sustained sounds of musical instruments. In the mean time, the shape of the vocal tract that determines the sounds is a characteristic of human articulator system, intensely studied in speech recognition tasks. Because singers have to sustain vowels as long as possible, they learn to develop a control technique over the pronunciation of the vowels (Barnes et al., 2004), thus putting difficulties in the use of techniques and models from speech processing (Youngmoo, 2003).

The cepstral coefficients are a set of features reported to be robust in some different pattern recognition tasks concerning human voice. They are widely used in speech recognition and also in speaker identification. Lately, research on musical instrument identification techniques proved the cepstral coefficients to be a useful set of features in this task also. The human voice is very well adapted to the ear sensitivity, most of the energy developed in speech being comprised in the lower frequency energy spectrum, below 4 kHz. In speech recognition tasks, usually the first 12 coefficients are retained, considering that they represent the slow variations of the spectrum of the signal (Rabiner and Juang, 1993), characterizing the vocal tract shape, the spectrum of the uttered words.

Attempts of using the same features in speaker recognition had proved that also identity features are coded into the cepstral coefficient representation of a sound. Experiments conducted on different number of speakers, with the use of neural networks in the modeling of categories and in the identification stage showed satisfactory results using a number of 12-14 coefficients (Seddik et al., 2004; Fredrickson and Tarassenko, 1995; Mafra and Simoes, 2004). Cepstral coefficients were also successfully used in instrument recognition: the use of 18 cepstral coefficients derived from a constant Q transform gives a good discrimination rate between oboe and clarinet (Brown, 1999), and the combination with temporal features can result in good instrument classification results (Eronen and Klapuri, 2000).

In this paper, a study of Mel-Frequency Cepstral Coefficients is proposed, concerning the identification of singing voices. In speaker identification systems, the low order coefficients were used, comprising vocal tract frequency information. The singing voice has a much larger variability than speech and much higher frequency components, starting with pitch, that can be up until 1200 Hz in soprano voices. The aim in this study is to determine if it is appropriate to characterize the singing voices using higher order cepstral coefficients, that are related to pitch and fine spectral structure rather than to the formantic structure. We try to determine if the lower or the upper subset of MFCCs encodes more individuality-related information.

The paper is organized as it follows: first we present a short review of the methods and processing for obtaining the cepstral coefficients, in section 2. Subsection 2.2 presents some basic concepts about neural networks and also the steps used in implementing, training and testing on different subsets of data. Section 3 will describe the study material, grouping of the data and training of the networks, and finally section 4 will present the results obtained for different network complexities in each case, giving the possibility to generalize the posed problem.

2 SIGNAL PROCESSING METHODS AND TOOLS

2.1 The Mel-Frequency Cepstral Coefficients

The cepstrum of a time domain signal $s(n)$ is the Inverse Fourier Transform of the log-magnitude spectrum of the signal. The log-magnitude spectrum of a real signal is a real and even function, thus the cepstrum is normally computed via Discrete Cosine Transform which is equivalent with the Fourier transform in case of even functions.

An important preprocessing step in the analysis of speech signals is the pre-emphasis of high frequencies. This is done because the amount of energy carried in the high energy components is small compared to low frequencies. For the singing voice, the high frequency components are all the more important for the perceived quality. Preemphasis is usually done by filtering the signal with a FIR filter whose transfer function in time domain is:

$$y(n) = x(n) - ax(n-1) \tag{1}$$

where a is close to 1, with typical values around 0.95.

The processing continues with a Fourier analysis of the windowed signal. A Hamming window of 20 ms was considered. The Mel-frequency scaling is done by a bank of triangular band-pass filters, nonuniformly distributed along the frequency axis. The Mel-scale equivalent value for frequency f expressed in Hz is:

$$mel(f) = 2595 log_{10}(1 + \frac{f}{700}) \tag{2}$$

The MFCCs are computed by redistributing the linearly-spaced bins of the log-magnitude FFT into Mel-spaced bins according to eq. 2, and applying DCT on the redistributed spectrum. A relatively small number of coefficients (typically 13) provide a smoothed version of the spectral envelope, leading to the isolation of the vocal tract response by the simple retention of the desired amount of information. An additional advantage in using MFCCs is that they have a decorrelating effect on the spectral data, maximizing the variance of the coefficients, similar to the effect of Principal Component Analysis. This allows the elimination of one of the preprocessing steps in the neural network training, which is the actual PCA to eliminate data redundancy.

2.2 Feed-Forward Neural Networks

The simplest architecture of a neural network is the feed-forward network, consisting of one or more hidden layers through which the signal travels one way only, from the input to output. This architecture is extensively used in pattern recognition because of its basic task of associating inputs with outputs. Properly trained backpropagation networks are able to generalize problems and to handle reasonably inputs they have never seen.

For improving the generalization of the neural networks during training and not get to the situation of overfitting the data, the early stopping method was used. The data set is divided into three subsets: a training set which will be used in training, a validation set and a test set. The error on the validation set is monitored during the training process; when the network begins to overfit the data, the error on the validation set will tend to rise, and the training will be stopped. This leads to a much faster training of the network, as long as we take upon the error, which will be larger than the imposed goal.

To improve the training of a network, certain preprocessing techniques can be performed. The one used in this study is normalization of mean and standard deviation of the training set so that the training and the target sets will have zero mean and unity standard deviation. The MFCCs are decorrelated and there is no need to check for data redundancy with PCA. Post-training analysis is used to check the performance of the trained networks.

3 SETTING UP THE EXPERIMENTS

3.1 The Database

The studied material consists in a number of 20 untrained voices. For each voice, there are two common musical phrases of medium length 3 seconds and a third different one of medium length 4 seconds, all sampled at 44100 Hz. The two common phrases were used as training data, and the third one for testing. It should be noted that while the models are constructed based on the same utterance, the identification uses different phrases for all the subjects. Four groups consisting of five voices were set up for initial experiments concerning identity characterization, and one group containing 10 voices was used to test the capabilities of neural networks to model the data in case of extending the database.

3.2 The Feature Set

The voice signals were pre-emphasized using a FIR filter as presented in eq. 1, with $a = 0.95$. MFCCs were calculated using the described method, as the DCT of the log-magnitude spectrum with 1024-point FFT. 32 MFCCs were calculated for each frame of the signal. The coefficients were partitioned for two different situations: coefficients 1–15 that characterize the smoothed spectrum, and coefficients 15–32 for the fine structure of the spectrum. The two subsets represent the input for training the neural network. Neural networks were trained also with the entire set of cepstral coefficients to check if any improvement is obtained by using all the available information.

3.3 The Neural Networks

For the groups of five voices, the neural network was chosen to have one 20-neurons hidden layer. One of the five neurons in the output layer was assigned to each voice by giving a positive unity answer. The number of neurons in the hidden layer was increased to 40 for modeling the group of 10 voices. We chose a training function that uses a variable learning rate set to 0.09 and with early stopping method. The validation data for this task was 1/4 of the whole training set. Initial experiments showed that neurons with tan-sigmoid transfer function perform much better in this recognition task than neurons with log-sigmoid transfer function. The input data was normalized so that all the coefficients have zero mean and unity variance. Using each subset of MFCCs, several networks were trained to ensure that we obtain the best results.

4 TRAINING RESULTS AND SIMULATIONS

For each group of five voices, a neural network was trained with the two common phrases. The early stopping method implies monitoring the error on the validation set during training. At first, the error will decrease, in the data fitting process, but in case the network starts to overfit the data, the error will rise and the training will return the weights from the minimum attended error. Usually the training stopped around 0.15 to 0.20 error, depending on the difficulty of modeling the data. The closer the value is to 1, the better the data fit for the corresponding voice. Based on the fit values we would expect best results in identification with the whole set of MFCCs. In the conditions of these results, we test the network with unknown data. The test phrase was processed through the same steps in order to obtain the sets of MFCCs and the coefficients were presented frame by frame as input to the trained network. We emphasize the fact that the test data is different for each voice, so in some cases it might resemble to the training data, while in others it can be very different. Table 1 summarizes the percent of correctly labeled frames and the degree of data fit for one group of five voices.

Although the correlation test shows better modeling of classes with the entire set of coefficients, it is not always necessary to use them all. Some voices can be distinguished by using the first 15 cepstral coefficients, while for others, the information in the upper coefficients gives the difference. In the mean time, using all of the coefficients in the same classification does not always provide a more reliable result.

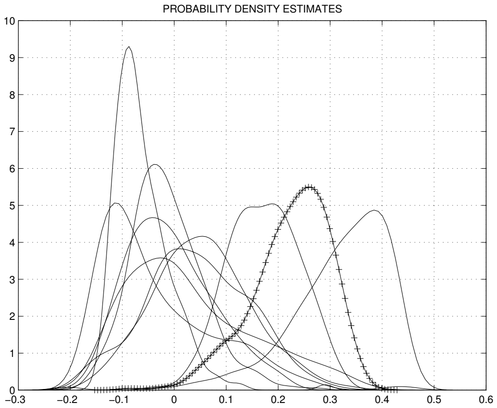

Figure 1: Probability distribution estimates of neurons responses for the test phrase; modeling with MFCCs 1-15

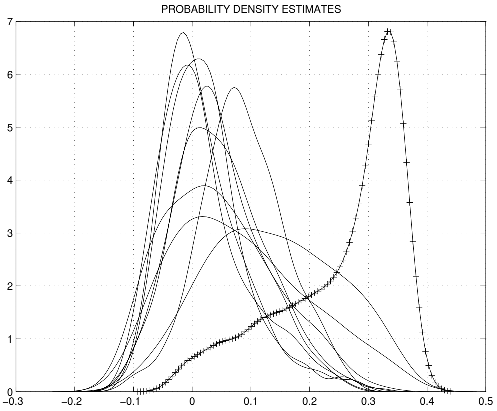

Figure 2: Probability distribution estimates of neurons responses for the test phrase; modeling with MFCCs 15-32

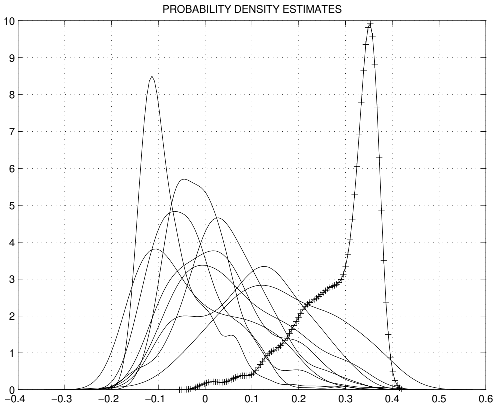

Figure 3: Probability distribution estimates of neurons responses for the test phrase; modeling with MFCCs 1-32

For increasing the number of voices used in the study, we trained a neural network with one 40-neurons hidden layer, using a set of ten voices, in the same conditions. Probability density estimates can be constructed based on the response of each neuron in each frame. The positive response neuron for one voice should have a PDE with mean close to 1, while the rest of the neurons should have PDEs close to 0. Figures 1-3 illustrate PDEs of the responses of the 10 neurons for the test phrase, estimated in 100 equidistant points, in one case that cannot be solved using low order MFCCs. The positive response neuron is represented by the '+' line. The generalization of the results state that the upper order cepstral coefficients contain at least the same quantity of information as the lower order ones. The cepstrum decomposes the problem in resonance-related information (low-order coefficients) and source-related information (high-order coefficients). As expected, both have their contribution to defining the identity of a voice, in singing, thus the source-related coefficients can be used to characterize the identity of the voice, and seem to behave well to changing the singing

Table 1: Correlation coefficient between target and network output for the training set and identification percent on the test phrase for one 5-categories experiment

	v01		v02		v03		v04		v05	
	fit	identif	fit	identif	fit	identif	fit	identif	fit	identif
coeff 1-15	0.69	0.46	0.67	0.79	0.63	0.60	0.60	0.81	0.62	0.47
coeff 15-32	0.68	0.53	0.59	0.63	0.65	0.64	0.61	0.64	0.56	0.47
coeff 1-32	0.78	0.50	0.78	0.80	0.78	0.74	0.79	0.84	0.79	0.54

style. Compared with the results obtained on speaker identification, it can be argued that in speech the filter part of the system does not have such a great variability as in singing, that is why the use of upper coefficients was generally not considered.

5 CONCLUSIONS

This paper presented a study of Mel-frequency cepstral coefficients in the context related to singing voice identification. The human articulator system in voicing is modeled in signal processing as a system with a specific signal - the glottal wave - as input to a linear time-invariant filter - the vocal tract. The low order cepstral coefficients represent information about the vocal tract shape, and the high order coefficients characterize the source signal. Both parts contain important information about voice identity.

In the case of singing voice, the input of the system is more invariant than the filter part. Cases difficult to handle with low-order MFCCs can eventually be solved correctly by using the high-order MFCCs. In this study no special care was taken for best trained neural networks; the purpose was rough and fast training for testing the selected features. For reliable results with neural networks in case of working with a large number of classes, parallel networks are used in order to achieve low complexity, fast training and small error rates in training each network.

6 FUTURE WORK

The results of the study lead to searching for a different way of characterizing the source in the articulator system, independently of the vocal tract parameters. A widely used method for estimating the glottal flow is through the Liljencrantz-Fant model; the processing involves determination of the closed glottis period, for correct inverse filtering. In singing and in high-pitched voices this is a real problem, because the closed glottis period may be too short for correct estimation of the inverse filter parameters. Also, authors of such studies used the voice signal and the simultaneous electroglottograph signal in order to locate specific instants in the voice signal. This method is inappropriate outside of laboratories, that is why we aim for an equivalent method of describing the glottal wave characteristics using information extracted only from the signal.

References

J. Barnes, P. Davis, J. Oates, and J. Chapman. The relationship between professional operatic soprano voice and high range spectral energy. *The Journal of the Acoustical Society of America*, 116(1):530–538, July 2004.

J. C. Brown. Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. *The Journal of the Acoustical Society of America*, 105:1933–1941, 1999.

A. Eronen and A.; Klapuri. Musical instrument recognition using cepstral coefficients and temporal features. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2:753–756, June 2000.

S.E. Fredrickson and L. Tarassenko. Text-independent speaker recognition using neural network techniques. *Fourth International Conference on Artificial Neural Networks*, pages 13–18, June 1995.

S. Hayakawa and F. Itakura. Text-dependent speaker recognition using the information in the higher frequency band. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:137–140, April 1994.

A.T. Mafra and M.G. Simoes. Text independent automatic speaker recognition using selforganizing maps. *39th IAS Annual Meeting Conference Record of the Industry Applications Conference*, 3:1503–1510, October 2004.

L. Rabiner and B-H. Juang. *Fundamentals of speech recognition*. PTR Prentice Hall, Englewood Cliffs, New Jersey, 1993.

H. Seddik, A. Rahmouni, and M. Sayadi. Text independent speaker recognition using the mel frequency cepstral coefficients and a neural network classifier. *First International Symposium on Control, Communications and Signal Processing*, pages 631–634, 2004.

F. Sun, B. Li, and H. Chi. Some key factors in speaker recognition using neural networks approach. *IEEE International Joint Conference on Neural Networks*, 3: 2752–2756, November 1991.

J Sundberg. Research on the singing voice in retrospect. *TMH-QPSR Speech, Music and Hearing*, 45:11–22, 2003.

E.K. Youngmoo. *Singing voice analysis/synthesis*. PhD thesis, Massachusetts institute of Technology, 2003.