Technical University of Cluj-Napoca
Faculty of Electronics, Telecommunications and Information Technology

Research regarding the modeling of musical signals timbre

ABSTRACT

Author: Annamaria Mesaroş
Scientific advisor: Prof. Dr. Eng. Corneliu Rusu
Cluj-Napoca 2007

Introduction

The musical timbre describes those characteristics that allow the human ear to distinguish between different sounds. The term *timbre* includes all the characteristics of a sound, except the frequency and the intensity of the sound, without giving an explicit list. Humans can tell the sound of a violin from the sound of a flute, the sound of one voice from the sound of another voice. The timbre allows the identification and tracking of the source of a sound, and humans recognize the musical instruments even without having any musical studies.

The singing voice is the oldest musical instrument. With the combination between music, lyrics and expression, the singing voice impresses us in a special way, compared to the other musical instruments. When singing voice is present, the listener is immediately attracted and focuses on the sound of the voice. This thesis proposes a study of the qualities and features that make the singing voice such a special instrument.

Individual voices are distinctive and reflect the identitiy of the singer. Once familiar with a singer's voice, we can easily identify him in other musical pieces. Our ability to recognize the singing voices is independent of the music itself. We are able to recognize the singer in musical pieces that we have never heard before. Also, we need very little amount of information before doing the identification. Sometimes one second or a phrase from the song is enough to get a compete idea of the definitory characteristics of the singer's voice [20].

A description of the distinctive character of a voice cannot be achieved without resuming to subjective terms as "harsh" or "squeeky", that do not have an objective correspondence. The qualities of the voice are a combination of physical factors like the size of the vocal tract and educated factors related to expresivity like for example the accent [16]. Quantizing, extracting and modeling these terms is a complicated problem [22]. The standard algorithms for the analysis and processing of the audio signals are not always appropriate to model the singing voice.

Similarly, understanding the perceptual features that make voice penetrate through the sounds of other musical instruments is difficult. Even identifying the presence of the voice in a mixture is difficult for a computational method, while humans do it instinctively. This difficulty extends over other classes of sounds. In a way, we know less about the perceptual features of the voice and more about modeling the vocal system.

The singing voice represents a challenge because of its large physical variation compared to the one of other instruments. To pronounce the words, one has to move the jaw, the tongue, changing the shape and the properties of the vocal mechanism [16]. This acoustical variation domain cannot be captured in a small order model. No other instrument develops the same quantity of physiscal variation like the human voice, and for this reason the audio processing techniques must be adapted to cope with the variation of this signal.

This thesis covers a study of the timbre of the musical instruments in order to characterize the identity of their source. Because of the interesting peculiarities of the singing voice, the study is directed towards the analysis of the singing voice and singer identification. The work is based on the assumption that the physical and expression features, as primary factors in determining the unique sound of a voice, can be represented by a number of features that would allow the distinction of the voices in a feature space using pattern classification techniques.

Organization of the thesis

The first phase of the study consists in the identification and extraction of the specific parameters. The parameters are estimated from singing voice recordings, using classical signal processing techniques. The second phase consists in modeling these parameters to capture the variation of the voice features, using automatic learning and classification algorithms.

Accordingly, the thesis is organized as follows. Chapter 2 presents basic information about the features of musical signals. This includes a general overview of the anatomy and physiology of the voice production system. The chapter also presents the main features of the musical instruments sounds and the correspondence between the objective. Physically measurable features and the subjective, perceptual features of musical signals.

Chapter 3 presents elements of musical signal processing. The chapter includes general methods for signal processing and specific methods that are used for voice processing. A detailed description of the methods for parameters estimation is given.

Chapter 4 presents general pattern classification methods that can be applied for classifying the singing voices based on the calculated features. The chapter introduces statistical signal processing terms and pattern classification techniques: discriminant functions, distances, neural networks, Gaussian mixtures. The reason for testing more classifiers and data sets for the same classification problem is also given.

Chapter 5 gives details about the author's contributions to this research area. It is developed in three directions: a study of the correlation between the time variation of spectral features, the study of the source-filter model for the voice production and separation of the two components and singing voice classification using Mel frequency cepstral coefficients as timbral features. The experiments are conducted to evaluate the different classification methods and proposed models, in the context of identifying a voice based on its numerical descriptors.

Chapter 6 is dedicated to conclusions and evaluation of the author's contribution.

Contributions

A major contribution of the thesis is the study of spectral characteristics of the singing voice. The correlation between the spectral features evolution in singing is the main subject of this part.

In order to achieve a complete characterization of the singing voice using numerical descriptors for features, we combine methods from speech analysis and musical signals analysis. The dynamics of the singing voice is different than the speech dynamics [16]. The parts with sustained vowels resemble to the sounds of the musical instruments. This thesis completes the study of the musical instruments sounds by adding the singing voice as a musical instrument, in the problem of determining a numerical description of its spectrum. Section 2 of the 5th chapter presents the estimation of the spectral features of the singing voice. The study uses voices singing the same phrase. The scope of this study is to observe and try to explain the dependencies beween the spectral features of the singing voice by combining knowledge from signal processing, musical theory and voice training.

The results indicate a strong dependency between the analyzed spectral features and the articulation, determined by the formants positioning. Professional singers train their singing voice to be able to shift the first formant frequency. In soprano voices the first formant is positioned at least at the value of the fundamental frequency, which is usually above the normal formant frequency value in speech. The voice quality is determined by the higher order formants, because the first two formants have predetermined positions for intelligibility. The position of the first formant influences all the studied spectral features. In speech, the formants position is given by the shape of the vocal tract and determines the uttered vowel, while in musical instruments the resonances lock at the multiples of the fundamental frequency [19].

The formants are characteristics of the vocal trcat shape and its frequency response. Two different components are involved in the voice production mechanism: the glottal wave as a source signal and the vocal tract as a filter which shapes the source signal to output the voice signal. The shape of the glottal wave contains information about the speaker's health and identity. Obtaining the glottal wave from the acoustic signal is theoretically achievable by estimating the vocal tract filter and filtering of the speech signal using the inverse filter. The vocal tract is modeled by an all-pole filter whose parameters can be obtained by linear prediction. The estimation of the system characteristics in the closed glottal phase gives an accurate model for inverse filtering of the voce signal [14], [21].

A contribution of this thesis to the study of source-filter model of speech production is presented in section 3 of chapter 5. In this area, the work proposes two methods for determining the glottal closed phase using only the acoustic signal.

The first step in determining the closed phase is locating the moments when the glottis closes - the glottal closure instant. For detecting the glottal closure instant, two methods given by different authors are used, namely using the group delay function and the Frobenius norm of the vocal signal covariance matrix [1], [4], [22].

The proposed method for determining the closed glottal phase is based on the formants frequencies. In the closed phase the signal is generated by free resonances of a tube, the frequencies of the formants are stable and they have large amplitude. In the open phase the system changes its characteristics due to the nonlinear coupling, the formants can have important jumps [1]. These changes are visible especially in the first and third formant frequency [15].

The vocal tract is modeled as a linear time invariant system with an all-pole transfer function. To determine the transfer function, an autoregressive model is used on a small size window to calculate the poles corresponding to the vocal tract resonances and determine the formant frequencies. By shifting the analysis window with one sample and repeating the formant frequency calculations, we get the evolution of the formants during the glottal cycle. The closed phase can be estimated as the period where the format frequency is almost constant.

The method gives reliable results for the singing voice with a low pitch. The analysis window size must be chosen such that the window is positioned in the closed phase enough times to get a number of formant frequency values. For the high-pitched voice, the sampling of the signal and its representation may include too few samples in the closed phase, thus no windowed analysis is possible.

Based on some observations made in using the methods for glottal closure instant determination, we consider the glottal opening as a secondary excitation, strong enough to influence the Frobenius norm of the covariance matrix of the signal. If the matix is constructed such that it contains ony one significant excitation, either a glottal closure or a glottal opening, the computed Frobenius norm will be multipeaked with local maxima at the closure and opening of the glottis. This gives a straightforward method to determine the glottal closed phase. Unfortunately, the proposed methods cannot be run automatically, being dependent on the analyzed signal.

A third contribution of this work is the construction of singing voice identification systems. Different combinations of features are used, different sets of cepstral coefficients and different methods for the classification. The cepstral coefficients are a set of robust features, used in speaker and also in instrument identification [2], [3]. Usual speaker identification systems based on cepstral coefficients use the lower order cepstral coefficients. Because the singing voice has different dynamics, we propose a classification based on the higher order cepstral coefficients. Preliminary results on singing voice identification using the two sets of coefficients and neural networks proves that the voices can be identified based on each of the two sets. On this background, we propose the construction of a number of singer identification systems.

There are powerful classfication algorithms suitable for construction also singer identification systems. For this task the linear and quadratic discriminant functions, distance based nearest neighbor classification and Gaussian mixture models were used. The results for these methods vary between 50% and 100% correct identification rate. The generalization of the results is obtained by using a 4-fold cross validation procedure.

The most robust method is the probabilistic one, using Gaussian Mixture Models and Maximum Likelihood classification. The advantage of the Gaussian Mixture Models is the possibility of increasing the number of mixtures and consequently the number of parameters and thus obtaining better performance. Depending on the number of components, the system can achieve the maximum recognition rate of 100% for some configuration of the training/testing, eith an average performance of over 90%.

Representative papers of the author in this domain are [7], [5], [6], [8], [9], [10], [11], [12].

Summary

This thesis presents the study of the features of musical signals for sound source identification based on musical timbre. The studied signal is the singing voice and it was chosen due to the complexity of its spectral features. The thesis presents elements of human perception and voice production, as definitory elements of the processing methods. The specific analysis methods for the voice signal are thoroughly presented.

One contribution of this thesis to the research area is the study of the singing voice spectral features, compared with the musical instruments sounds and speech. There is an important dependency to notice, all the spectral features being dependent on the position of the first formant. In speech, the first two formants have predetermined positions needed for speech intelligibility, while in singing, the first formant is always adjusted to macth at least the pitch of singing, which can be very high. The tuning is achieved through musical training and experience.

The second contribution is drawn to the separation of the voice characteristics into the two components of the voice production system: the glottal wave and the vocal tract. To obtain the glottal wave by means of inverse filtering, the thesis proposes two methods for detecting the glottal closed phase. The first method is based on the first formant modulaton during one glottal cycle, the formant having a constant frequency in the glottal closed phase. The second method is based on the localization of the local maxima in the Frobenius norm of the covariance matrix of the analyzed signal.

The thesis also constructs singer identification systems for monophonic music. Solo voice recordings are used, with up to 100% correct voice identification performance.

The thesis approaches a very interesting problem: what are the features that define the uniqueness of a singer's voice? The models and methods have just a limited success, limitations coming from the small size of the study set. With the database used in the work, the methods have a good performance to distinguish between different voices and to capture the important features that define the identity of the voice. The models and methods have a limited success though, with limitations coming from the small size of the database, which does not allow generalization.

Selected bibliography

[1] T.V. Ananthapadmanabha and B. Yegnanarayana. Epoch extraction from linear prediction residualfor identification of closed glottis interval. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-27(4), 1979.

[2] 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, 1999.

[3] A. Eronen and A.; Klapuri. Musical instrument recognition using cepstral coefficients and temporal features. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2, 2000.

[4] M.R. Matausek and V.S. Batalov. A new approach to the determination of the glottal waveform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-28(6), 1980.

[5] A. Mesaros. Modelarea individualitatii vocii cantate prin coeficienti cepstrali si retele neuronale. *Workshop Verificatori Biometrici*, 2005.

[6] A. Mesaros. Estimation of closed glottis phase in professional singing voice using the frobenius norm. *Analysis of Biomedical Signals and Images, Proceedings of 18th Biennial International EURASIP Conference Biosignal 2006*, 2006.

[7] A. Mesaros. Spectrum characteristics of singing voice signals and their usefulness in singer identification. *6th Communications International Conference, COMM2006*, 2006.

[8] A. Mesaros and J. Astola. Inter-dependence of spectral measures for the singing voice. In *International Symposium on Signals, Circuits and Systems*, Iasi, Romania, 2005.

[9] A. Mesaros and J. Astola. The mel-frequency cepstral coefficients in the context of singing voice. *International Conference on Music Information Retrieval*, 2005.

[10] A. Mesaros and E Lupu. Closed phase detection in the singing voice using information about formant frequencies during one glottal cycle. In *Proceedings of 10th International Conference on Speech and Computer*, Patras, Greece, 2005.

[11] A. Mesaros and S. Moldovan. Methods for singing voice identification using energy coefficients as features. *2006 IEEE-TTTC International Conference on Automation, Quality and Testing, Robotics AQTR 2006 (THETA 15)*, 2006.

[12] A. Mesaros and S. Moldovan. Methods for singing voice identification using energy coefficients as features, acceptată pentru publicare. *Acta Technica Napocensis*, 48/3, 2007.

[13] I. Nafornita, A. Campeanu, and A. Isar. *Semnale, circuite si sisteme, partea I.* Universitatea Politehnica Timisoara, 1995.

[14] M. Rothenberg. Research aspects of singing. *Royal Swedish Academy of Music*, 1981.

[15] J Sundberg. Research on the singing voice in retrospect. *TMH-QPSR Speech, Music and Hearing*, 45, 2003.

[16] Johan Sundberg. *The Science of the Singing Voice.* Northern Illinois University Press, 1987.

[17] D Tarniceriu. *Bazele prelucrarii numerice a semnalelor.* Vasiliana, Iasi, 2001.

[18] G. Toderean and A. Caruntu. *Metode de recunoastere a vorbirii.* Risoprint, Cluj-Napoca, 2005.

[19] Dem Urma. *Acustica si muzica.* Editura Stiintifica si Enciclopedica, 1982.

[20] G. H. Wakefield and M. A. Bartsch. Where's Caruso? Singer identification by listener and machine. In *Cambridge University Music Processing Colloquium*, 2003.

[21] B. Yegnanarayana and H. A. Murthy. Significance of group delay functions in spectrum estimation. *IEEE Transactions on Signal Processing*, 40(9), 1992.

[22] B. Yegnanarayana and R.N.J. Veldhuis. Extraction of vocal tract system characteristics from speech signals. *IEEE Transactions on Speech and Audio Processing*, 6(4), 1998.