Towards the Application of Fuzzy Logic to the Sound Recordings Fingerprint Computation and Comparison

LADISLAVA JANKU, LENKA LHOTSKA
Department of Cybernetics, Faculty of Electrical Engineering
Czech Technical University in Prague
Technicka 2, 166 27 Praha 6
CZECH REPUBLIC

Abstract: This paper deals with the application of computational auditory scene analysis system and voice identification algorithms in audio fingerprinting. Generally, audio fingerprinting technology generates a unique fingerprint for an audio file based on an analysis of the acoustic properties of the audio itself and represent unique identification key of the music piece. Presented approach enables extract voice features and compute the hash function, which depends on these voice properties. A system for voice extraction through exploration of the auditory scene analysis approaches and fuzzy rule system is presented.

Key-Words: fuzzy logic, computational auditory scene analysis, audio fingerprint

1 Introduction
This study is motivated by an interest in the properties of sound signals and identification of them with the practical application of automatic searching in large music databases containing music data in signal representation or in compressed format.

In recent years, audio fingerprinting technology was suggested as a sufficient and powerful tool for audio sample identification. Audio fingerprint is a unique key of each sound sample. This property brings about the following effect – a music piece, performed by one performer, is characterized by the different audio fingerprints that the same music piece performed by another musician. This property of audio fingerprinting causes a limitation of searching facilities. It’s not possible to use them for melody-oriented or performer-oriented search.

This weakness can be partially avoided by extension of the audio fingerprint generation algorithm. Approach presented in this paper involves physiologically motivated sound pre-processing, voice feature extraction and hashing. It combines model of auditory periphery with audio features extracting mechanism, and fuzzy rule system.

2 Audio Fingerprints - Background
Audio fingerprinting technology generates a unique fingerprint for an audio file based on an analysis of the acoustic properties of the audio itself. Each audio fingerprint is unique, e.g each sound has a unique audiofingerprint.

Audio fingerprint is generated by the application both of acoustical features extracting algorithms and by one-way hash functions. Figure 1 shows the whole process of the audio fingerprint generation. Input sample is divided into 30-70 narrow frequency bands.

![Fig 1: Standard audio fingerprint generation method](image)
creation. Some systems use the energy levels and energy distribution in each of the narrow frequency bands. For example, the fingerprint technology of Songprint [1] explores following algorithms: computing FFT, dividing sample in 64 narrow frequency bands, extracting energy envelope in each frequency band, extracting non-zero parts, computing average and standard deviation over the frequency bands.

Audio fingerprints are very small. They can be used to the identification of unknown music pieces.

Such audio identification system converts input sound into the hash value (audio fingerprint) and matches this audio fingerprint with the audio fingerprints within the database. Figure 2 shows a whole process of the unknown music piece identification. User sends a request for the music piece identification to the system.

![Fig. 2: Music piece identification using audio fingerprint](image)

This system computes a hash code of this music piece and tries to match it with the hash codes in the database. If the computed fingerprint matches some fingerprint in the database, the identification parameters included in the matching record are sent back to the user.

2.1 One-Way Hash Functions

One-way hash function is a function, which operates on arbitrary length input and returns a fixed length hash value (code). In addition, these functions have some other characteristics that make them one-way.

It can be formulated mathematically in the following way:

\[ c = H(A) \]

where \( c \) represents a generated code (hash value), \( H \) is a hash function and \( A \) represents arbitrary-length input sequence. A hash function has the following features [2]:

- Given \( A \), it is easy to compute \( c \).
- Given \( c \), it is hard to compute \( A \) such that \( H(A) = c \).
- Given \( A \), it is hard to find another input sequence \( A' \), such that \( H(A) = H(A') \).

One-way functions are built on the idea of compression function. This one-way function outputs a hash value of length \( n \) given an input of some larger length \( m \). A lot of one-way hash functions were designed. Snefru, N-hash, MD4, MD5, SHA, HAVAL belong among the most known hash functions. The audio fingerprinting algorithm use usually their own hash functions, but these hash functions are based on the hashing principles of the hash functions listed above.

3 Auditory Scene Analysis

One of the main goals of research in computational auditory scene analysis is to create computer systems that can separate and recognize sound source in a complex auditory environment. In contrast to commonly used speech recognition methods consisting in statistical models of sound generation [4], we follow machine listening approach, which consists in modelling the detection mechanism (human auditory processes).

Auditory scene analysis covers the entire hearing process beginning with the reception of sound signal, through several stages of auditory processing, and culminating in the formation and separation of auditory sources [3]. This process is complex and poorly understood. It occurs in two conceptually distinct stages. In the first stage, sound is decomposed into a collection of sensory elements. In the second stage, elements that are likely to have arisen from the same sound source are grouped to form a perceptual whole [15].

Computational auditory scene analysis covers the area of systems based on principles of physiologically inspired grouping heuristics and learned schemata described by Bregman [9]. Numerous approaches to the problem of source separation have been developed. Early work on audio separation was done by Stockham, who used homomorphic signal processing to separate Caruso’s voice from background noise and accompaniment in recordings made in 1908. Stockham was a pioneer, but the real research in the area of sound source separation started in the middle of 1980’s. Schneider and von der Malsburg’s neural network model [10], Weintraub’s state-dependent model, Brown and Cooke’s data-driven model [5, 6], Ellis prediction-driven model [7] implemented on the basis of the blackboard architecture of the IPUS system, Nakatami and Okuno’s multi-agent system models for auditory scene analysis [11], or another approach based on oscillatory correlation [14].

4 Problem Formulation

Audio fingerprinting technology enables to identify uniquely a sample and can be used for searching large multi-media databases containing music data in signal representation or in compressed format. In the other
hand, user needs for such a kind of searching process a part of the searched sample, music piece, etc.

Audio fingerprinting technology doesn’t enable to explore other useful searching criteria - melody, performer identification, music instruments identification, etc. Searching facilities exploring these techniques could offer to each user a wide range of features, for instance searching of music pieces containing sounds from user specified musical instruments. For this reason, these searching systems could be very user-friendly. Another application of musical instruments classification is an application in the line of audio indexing.

Computing a performer audio fingerprint independent on the performed music piece requires extracting such audio features from the input sample, which are essential for the music performer identification. In melody dependent searching, the situation is very similar, only the melody extracting algorithm is needed. The development of this audio fingerprint facility was inspired through this idea.

5 System Structure

The unique performer/singer identification technology sets a new requirement to whole system. The process of audio sample fingerprint computation is match more complex, and required more sophisticated algorithms, particularly in the area of singer’s voice features extraction. The basic system structure is shown on Figure 3. Whole process consists of three phases:

- voice extraction
- singer dependent features extraction
- hashing.

The voice extraction phase is realized through the application of computational auditory scene analysis system. Our approach to auditory scene analysis system reflects the need of uncertainty expression and explores all advantages of fuzzy logic and fuzzy rule based systems. The voice extraction system architecture description constitutes a main topic of this paper.

Generally speaking, the task of performer’s voice features extraction is very similar to the task validation of the identity of a person by his her voice only. Prior to verification, the user must be enrolled onto the system so that a model of his her voice can be created. Classification methods can be grouped together in to two groups: statistical methods that include Gaussian Mixture Models [16] and discriminant methods, which include multilayer perceptrons and polynomial classifiers. Last phase consists in application of an appropriate hash function.

6 Voice Extraction through Models of Auditory Processes and Fuzzy-Rule System

The auditory scene analysis system consists of several stages. In the first stage, peripheral auditory processing is simulated by a bank of bandpass filters and a model of inner hair transduction. In the second stage, following features are extracted for each cochlear channel – amplitude modulation, onsets, time envelope, pitch and energy shifts in time domain. In the third stage, incoming features are compared with the features predicted in the previous step. If they are matched, they are grouped both on basis of the knowledge of real sound structure and on basis of the real actual world model. If not, they are grouped on basis of the knowledge of sound structure only. Possible real-world models are constructed and evaluated by a range of criteria. This evaluation is not a simple process. Each criterion corresponds to some heuristic derived from the characters of the natural sound scene or to some auditory grouping heuristic. A problem of sound scene analysis is regarded as a multiple criteria decision problem. The aim is to select the best matching real-world model from the set of possible real-world models.

The grouping mechanism is supplied by the decision logic evaluating possible real-world models by applying fuzzy rules. A problem of sound scene analysis is regarded as a multiple criteria decision problem. Note, that possible real sound scene models (real-world models) are computed for each time element of the input representation. This computation is based on the features extracted from the data obtained in some time interval before the time, in which is computed and on prediction of the set of the features for the next time element.
As we said above, possible real-world models are evaluated by a set of criteria. Each criterion corresponds to some heuristic derived from the characters of the natural sound scene or to some auditory grouping heuristic. For these criteria, see our previous work. The fuzzy preference relations are constructed to express relations between possible real-world models [17]. Particular attention is given to the application of strict preference, incomparability and non-dominated preference relations. Constructed fuzzy relations are aggregated by the application of the OWA operators [17, 18].

7 Conclusion
Audio fingerprinting technology generates a unique fingerprint for an audio file based, which represents unique identification key of the music piece. Presented approach enables extract voice features and compute the hash function, which depends on these voice properties.

A system for voice extraction through exploration of the auditory scene analysis approaches and fuzzy rule system is presented. This paper describes a simple system performing a task of voice separation. In conclusion, we give a short overview of this contribution. A detail description, on how developed system works, is given. The core of this system represents fuzzy rule system. Only essential auditory features are extracted. More sophisticated artificial recognition systems would probably need wider range of features (spectral envelope, spectral centroid, inharmonicity, etc.) or high-level features (musical instrument identification, etc.).

Possible real-world models are evaluated in the following way - our solution is based on assumption that the problem of auditory scene analysis can be regarded as a problem of the multiple criteria decision problem. The sound source high-level features can improve a signal separation of each sound source, which embodies any long-time characteristic (rhythm, etc.), which produces sound textures, which moves (moving sound source trajectory prediction), or whose recognition can be improved by learning (specific sound of any musical instrument, speaker identification). Naturally, high-level features are related to the corresponding low-level features. Fuzzy rule system represents a core of this system. Fuzzy preference relations are constructed and aggregated by the application of the OWA operators.

All computations mentioned in this work are expensive but most of them are easily adapted to parallel processing architectures.

Acknowledgements
This research work has been supported by the research program Transdisciplinary Biomedical Engineering Research MŠMT:21000012 of the Czech Technical University in Prague (sponsored by the Ministry of Education, Youth and Sports of the Czech Republic).

References: