Audio Interaction with Multimedia Information

MARIO MALCANGI
DICo – Dipartimento di Informatica e Comunicazione
Università degli Studi di Milano
Via Comelico 39 – 20135 Milano
ITALY
malcangi@dico.unimi.it

Abstract: - Interacting with multimedia information stored in systems or on the web points up several difficulties inherent in the signal nature of such information. These difficulties are especially evident when palmtop devices are used for such purposes. Developing and integrating a set of algorithms designed for extracting audio information is a primary step toward providing user-friendly access to multimedia information and developing powerful communication interfaces. Audio has several advantages over other communication media. These include: hands-free operation; unattended interaction; simple, cheap devices for capture and playback. A set of algorithms and processes for extracting semantic and syntactic information from audio signals, including voice, was defined. The extracted information was used to access information in multimedia databases, as well as to index it. More extensive, higher-level information, such as audio-source identification (speaker identification) and genre (in the case of music), must be extracted from the audio signal. One basic task involves transforming audio into symbols (e.g. music transformed into a score, speech transformed into text) and transcribing symbols into audio (e.g. score transformed into musical audio, text transformed into speech). The purpose is to search for and access any kind of multimedia information by means of audio. To attain these results, digital audio processing, digital speech processing, and soft-computing methods need to be integrated.

Key-Words: - Audio features, multimedia information, speech-to-text, audio-to-score, text-to-speech, score-to-audio, digital audio processing, pattern matching, softcomputing.

1 Introduction
The traditional approach to searching for information in storage systems (databases) and networks (webpage content) is limited to text. Engines such as Google are able to search for information where an alphanumeric string matches webpage content. Multimedia information can be retrieved only if text is embedded in it. Such text is not exhaustive of multimedia information. Therefore, much of the multimedia information will not be available to search engines unless a preprocessing action, such as indexing or transcribing text (titles), has been executed.

Much multimedia information cannot be precisely represented by text because most such information is not strictly semantic [2][15]. Audio and video are very rich in information content, but audio is part of the video so that video information is also related to audio. Videos can be indexed by means of audio fragments more exhaustively than by text indexing. For this reason, a primary step in building a multimedia search engine must focus on audio classification.

Audio is the component of multimedia information characterized by a certain peculiarity. It is simple to capture (microphones are very cheap, always available in most embedded devices, very simple to use in unattended mode, etc.), does not require a high sampling rate to be digitized, and is relatively easy to synthesize.

Audio exists in three main forms, depending on its source: voice, music, and generic sounds. Each of these is a very different form of audio information, showing to what extent audio can richer in information than any other medium. This is also a disadvantage in terms of processing, because different algorithms need to be applied for pattern-matching and synthesis purposes.

A great deal of research has concerned speech processing (recognition, identification, synthesis, encoding, and decoding) [1][16][17]. Music has also received researchers’ attention for several decades, particularly regarding synthesis [13][14][6]. Only recently has much attention focused on genre recognition and on stream retrieval.

Less attention has been aimed at general purpose sounds, primarily because such sounds are generated by different sources and carry different information. Such audio is no less information-important than speech or music. An example is speech mixed with sounds such as hiccups, coughs, and so forth or (modern) music mixed with generic sounds.

Speech-to-text and music-to-score are two basic strategies for empowering a traditional search engine that matches text and/or string sequences on the basis of traditional text-retrieval tactics. When text and scores are not available in a multimedia database, pattern matching is the only strategy that can lead to success for an application that provides interaction and access to multimedia information.
Applying pattern matching to audio is a very complex task. It consists of two main processing subtasks, feature extraction and distance evaluation. Feature extraction refers to classical digital signal-processing algorithms. Distance evaluations can be approached through hard-computing methodologies, such as dynamic-time warping (DTW) or hidden Markov modeling (HMM) [9], or through soft-computing methodologies such as fuzzy logic (FL) [5][10][11][19][20] or artificial neural networks (ANNs) [4][7][8][12]. A combination of the above methodologies may also represent a good solution, especially when the pattern-matching task is very complex.

2 Audio-feature Extraction

Audio features are extracted from raw audio data in the time and frequency domain, considering short 50%-overlapped Hamming-windowed frames. A 20-millisecond frame duration is used for feature measurement, so that a set of feature-time sequences are generated.

The main time-domain features are RMS (root mean square) and ZCR (zero-crossing rate). Other time-domain features can be derived from the above through more complex processing so as to highlight certain specific audio properties.

RMS is the measurement of audio loudness. This feature is essential in tracing the presence or absence of audio or in characterizing an audio source by its dynamics:

$$RMS(n) = \sqrt{\frac{1}{N-1} \sum_{m=0}^{N-1} s^2(m)}$$

Zero-crossing rate is a time-domain-based frequency measurement. This feature is a good indicator of the nature of the audio frame, in terms of how much it may be pitched or unpitched.

$$ZCR(n) = \sum_{m=0}^{N-1} 0.5|\text{sign}(s(m)) - \text{sign}(s(m-1))| \omega(n-m)$$

The main frequency-domain features are pitch and band-spectrum. Other frequency-domain features can be derived from these to focus specific audio characterization, e.g. to distinguish between speech and music.

Pitch (P) is measured using an autocorrelation function:

$$AC(i) = \sum_{j=1}^{N} \sum_{j=1}^{N+1-i} x(j)x(i+j-1)$$

Band-Spectrum (BS) is computed using a short-term Fourier transform (STFT). The frequency spectrum is divided into bands representative of frequency grouping in audio (formants for speech, harmonic distribution for music, etc.):

$$BS(j) = \{0, \frac{\omega_1}{8}, B_2 = \{\frac{\omega_1}{8}, \frac{\omega_2}{4}, B_1 = \{\frac{\omega_1}{4}, \frac{\omega_2}{2}, B = \{\frac{\omega_2}{2}, 0 \}$$

Other features can be also computed using more complex computation, such as Cepstrum and Linear Prediction Coding (LPC). Such measurements can be useful to extend the ability of the pattern matcher to support special applications such as speaker identification or melody tracing.

3 Hard-computing Pattern Matching

Hard-computing pattern matching refers to two main methods, primarily for speech recognition: DTW and HMM.

The dynamic time-warping algorithm is used to align and compare audio templates with the audio segment to be recognized. The templates consist of a combination of coefficients extracted from each audio frame. The DTW algorithm computes the matching cost for each audio item in the set of searchable audio patterns.

Euclidean distance is used to score the audio patterns to be recognized. The Mahalanobis distance measurement

$$D_i(x) = (x - \bar{x})^T W^{-1} (x - \bar{x})$$

requires that the test pattern x be processed with reference to the averaged feature vector \( \bar{x} \) representing the audio item to be found. The distance \( D_i(x) \) is used as a score for the \( i \)-th audio item in the pool of searchable sounds.
4 Soft-computing Pattern Matching

Soft-computing pattern matching refers to two main methods: FL and ANNs. Each has its own advantages, so that choosing between them is a matter of the nature of the information to be matched.

4.1 Fuzzy logic-based pattern matching

Fuzzy logic is a novel approach to pattern matching, as it was traditionally and successfully used in control applications. Only in recent years has part of the research focused on exploring fuzzy logic’s ability to match patterns.

A membership function was defined for each measured feature to transform it from the crisp domain to the fuzzy domain. The normalized measurements are fuzzified according to appropriate membership functions.

A set of rules was defined to model the way a specific class of audio patterns can be recognized by a human being. An expert in audio perception transfers her or his knowledge into these rules, thus tuning the FL engine to recognize such sounds as the human expert does. A typical rule has the following format:

IF \( f_1 \) IS (fuzzy value) AND IF \( f_2 \) IS (fuzzy value) AND … THEN the item IS (fuzzy value)

4.1 Pattern-matching artificial neural networks

Artificial neural networks (ANNs) can be very useful in matching an isolated audio pattern because of their ability to compensate for the time variation in the input vis-à-vis the template. A three-layer, feed-forward, back-propagation artificial neural network (FFBP-ANN) can be used for this purpose. Such an ANN has its inputs fully connected to all the nodes of the hidden layer. The hidden layer is also fully connected to the output nodes.

All the inputs have a linear activation function. A sigmoid activation function connects hidden-layer nodes to output-layer nodes.

A non-linear (sigmoid) activation function connects hidden-layer nodes to the output layer:

\[
s_i = \frac{1}{1 + e^{-I_i}}.
\]

\[
I_i = \sum_j w_{ij} s_j
\]

\( s \) is the output of the \( i \)-th unit
\( E_i \) is the total input
\( w_{ij} \) is the weight from the \( j \)-th to \( i \)-th unit

The ANN’s inputs are the binary-encoded audio features. Such inputs, which refer to an audio frame, can be sequenced so only stationary frames stimulate the ANN.

---

**Figure 2. Membership functions**

**Figure 3. Architecture of the FFPB-ANN**

The ANN’s outputs encode the audio frame corresponding to the features fed to its inputs. Isolated, as well as continuous, audio frames can be identified by the ANN, depending on its training.

5 Audio Spotting

Audio spotting can be considered an initial level of applying hard-computing and soft-computing methods to the challenge of accessing multimedia information by means of audio input.

Audio spotting is a search procedure whereby an audio pattern is continuously compared to an audio stream so as to find the frame that best matches what was searched for. This is a kind of continuous pattern matching.

Hard-computing methods, such as DTW and HMM, can be used for this purpose, but soft-computing methods prove more effective. This is due to the ability of the ANN to work like a window that moves along the signal.
References:


