December 2001

Fuzzy clustering evaluation of time-frequency distribution (TFD) schemes for audio stream segregation

M. A. Jackson
University of Wollongong

I. Burnett
University of Wollongong, ianb@uow.edu.au

Fuzzy clustering evaluation of time-frequency distribution (TFD) schemes for audio stream segregation

Abstract

Audio stream segregation is a task performed constantly by the human auditory system, yet is difficult to reproduce with a computer. The research detailed in this paper looks at performing just one method of stream segregation - the temporal coherence boundary - using a fuzzy clustering system. The main focus of the paper is on examining the effectiveness of several time-frequency distributions as the feature vectors for the system. Three time-frequency distributions are examined and their effectiveness evaluated in terms of correct separation and computational complexity. The main evaluation compares the popular gamma-tone filter bank with the MPEG-7 audio spectrum envelope. The results are promising, indicating that the less computationally expensive MPEG-7 descriptor performs well, implying that stream segregation may be possible using the MPEG-7 audio low-level description scheme.

Disciplines

Physical Sciences and Mathematics

Publication Details

This article was originally published as: Jackson, MA & Burnett, IS, Fuzzy clustering evaluation of time-frequency distribution (TFD) schemes for audio stream segregation, 10th IEEE International Conference on Fuzzy Systems, 2-5 December 2001, vol 2, 553-556. Copyright IEEE 2001.

Fuzzy Clustering Evaluation of Time-Frequency Distribution (TFD) Schemes for Audio Stream Segregation

Melanie A. Jackson, Ian S. Burnett
School of Electrical, Computer and Telecommunications Engineering
University of Wollongong
Northfields Avenue, Wollongong, NSW, 2522, Australia

Abstract

Audio stream segregation is a task performed constantly by the human auditory system, yet is difficult to reproduce with a computer. The research detailed in this paper looks at performing just one method of stream segregation, the temporal coherence boundary, using a fuzzy clustering system. The main focus of the paper is to examine the effectiveness of several time-frequency distributions as the feature vectors for the system. Three time-frequency distributions are examined and their effectiveness evaluated in terms of correct separation and computational complexity. The main evaluation compares the popular gamma-tone filterbank with the MPEG-7 Audio Spectrum Envelope. The results are promising indicating the less computationally expensive MPEG-7 descriptor performs well; implying stream segregation may be possible using the MPEG-7 Audio low-level description scheme.

I. Introduction

Auditory scene analysis as performed by the human auditory system has provided researchers with a puzzle for many years. The field of research known as Computational Auditory Scene Analysis (CASA) endeavours to replicate the performance of the human auditory system with a computer. Numerous researchers have applied different computational techniques to the scene analysis task. Some researchers use statistical methods to decompose the mixed audio signal into statistically independent sources [1][2], without exploiting the knowledge provided by the psychology field. Other groups use the psychoacoustic knowledge and methods such as blackboard architectures to develop hypotheses [3][4][5].

A fully working CASA system would provide a technique to label audio data with meaningful descriptions. This coincides with the MPEG-7 Standard, currently under development, a standardized multimedia description scheme. It includes a set of audio descriptors, which enable multimedia searching. Current techniques for labelling audio files with high-level descriptions are predominantly manual, minimal progress has been made in automatic transcription from low-level descriptors to high-level descriptors. CASA is one field of research that aims to fill the void.

Psychoacoustic studies have revealed patterns in the auditory system with respect to the formation of streams in mixed source environments. It is known that there are two levels of analysis performed in the human auditory system: innate processes and schema-based processes [6]. The schema-based processes are well developed especially the schema of spoken language. The innate processes elude computational techniques.

It is the innate processes that are explored in this work. In particular c-means fuzzy clustering is used to perform stream segregation based on the temporal coherence boundary [7]. The external processing completed in the ear canal and basilar membrane has been explored. It is understood that the basilar membrane provides a time frequency analysis of the signal to higher levels of the auditory system [8]. The gamma tone filterbank, is believed to be a good approximation to the motion of the basilar membrane [9]. However, the computational complexity of current implementations of this filtering makes the filterbank impractical for real-time application. The performance of the gamma tone TFD and two other methods for TFD calculation are evaluated here, as a pre-processor for the fuzzy clustering system. The performance characteristics include computational speed and effective clustering. The other transforms considered are a linear spectrogram and the MPEG-7 Audio Spectrum Envelope at varying resolutions.

Background on Audio Stream Segregation

Psychologists have identified a number of features that are used to perform stream segregation [6]. Stream segregation (or source separation) is the name applied to the allocation of acoustic energy to particular sources. Examples of human audio source separation include identification of two different speakers speaking at one time or identifying, in the everyday environment, events such as conversations, air conditioning whirs, computer keys clicking and/or background radio music. The human auditory system segregates the energy quite easily, using numerous perceptual cues, including common amplitude and amplitude modulation, temporal proximity, spectral proximity, harmonic relations, common onset, spatial cues and others. By playing simple acoustic patterns to subjects psychologists have shown relationships among these cues [6][7][10][11][12]. This work examines specifically the relationship between temporal and spectral proximity in simultaneous streaming.

Fuzzy Cluster Technique

Many of the researchers who have performed work in the CASA area apply strict allocation of TFD elements to single streams. Results have shown that the boundary of temporal coherence is not a strict boundary [7], hence, fuzzy clustering is an appropriate approach for dividing a TFD of a signal into streams of information.

Two streams may overlap in the TFD and hence partial allocation of energy in the overlapping regions to overlapping sources is required. This may also be achieved through fuzzy clustering.

The technique was originally explored by [13] arbitrarily using the gammatone filterbank to get a TFD and perform a fuzzy c-means (FCM) clustering on the TFD. The results are reasonable, however these do not indicate the scaling of the data dimensions required to achieve good clustering. Assuming the Euclidean vector norm is used in a FCM clustering system scaling of the time and frequency values is required to ensure the grouping corresponds to the known temporal coherence boundary information.

The difficulty with using FCM is predicting the required number of clusters. Subtractive clustering might be employed to predict the number of clusters required in a real auditory scene. However in the test input used here the number of clusters desired is known.

MPEG-7 Low-level Descriptors

The sound archives in the world, even those held digitally are a vast resource. Searching this audio data is restricted by the current meta-data associated with each file or by analyzing the raw signal data (or coded audio data). The difficulty arises when the meta-data insufficiently describes the file content. MPEG-7 provides a new

This work was supported in part by an Australian Postgraduate Award and a Motorola (Australia) Partnerships in Research grant, and Motorola, Inc., Australia.

standard for descriptive meta-data for multimedia files, including audio. The advantage of MPEG-7 is in the low-level descriptions, directly and automatically derivable from the raw signal.

The MPEG-7 standard, at the time of writing, is at final committee draft status. At present there are seventeen MPEG-7 Audio Low-level descriptors (LLDs). Each of the descriptors may be derived from a sampled waveform of an audio signal. Several of the LLDs allow several implementations to create a descriptor, for example the Audio Fundamental Frequency descriptor may be estimated using a number of different approaches. The importance is that these descriptors will be available as meta-data or can be easily generated from audio files.

The question we pose here is "How can these descriptors be used in CASA?" The intent of this research is to employ a set of Gestalt principles to implement a computational system to replicate innate auditory processing. Each of the MPEG-7 Audio LLDs needs to be examined to see how it may identify a Gestalt principle. For example the Audio Power descriptor may indicate the onset of a new source; the Audio Fundamental Frequency may indicate, or at least aid in the detection of, tonal components. This paper proposes the Audio Spectrum Envelope descriptor as a useful descriptor for temporal and spectral proximity.

II. Psychoacoustic Stream Segregation

Temporal Coherence and Fission

Van Noorden [7][11] thoroughly examined the ability to fuse the tones into a single coherent stream based on the amplitude difference and frequency separation of alternating tones. Stream segregation (fission) was evident when just one of these variations was present, either amplitude or frequency. The extent of the research performed on frequency separation and what is called the 'temporal coherence' boundary provides excellent parameters for scaling of the each dimension of the TFD of the signal.

Van Noorden examined how variation in the tone repetition rate, T, the tone duration, D, and the mean tone interval, Δf, changed the perception of alternating tones from a single stream to two steams. His audio data consisted of a series of tones of equal amplitude alternating in frequency, as illustrated in Fig. 1. Hence Van Noorden was able to determine the temporal coherence boundary. Figure 5.11 in [7] shows a general relation between the frequency separation (in semitones) and the time between the two tones adjacent to the tone in question (2T-D).

Even though Van Noorden's experimental data gave an exponentially increasing curve, for small 2T-D (<350ms) the curve is approximately linear. Taking an approximation to the slope in this region should give a reasonable relationship between the strength of temporal proximity and of spectral proximity. The approximation is shown in Fig. 3. The ratio is approximately 42.86 semitones per second. Hence if the features of an alternating sequence give a spectral to temporal ratio less than 42.86 semitones/s then temporal coherence occurs, if the ratio is found to be greater then fission will occur.

Bregmans Input

Our test input was a prepared set of alternating tones, f_a and f_b, with a capturing tone presented, f_c, synchronous to the spectrally closer of f_a and f_b, as illustrated in Fig. 2 [14]. The testing now varies from Van Noorden's testing, as the object is to separate the test data into two streams. The capturing tone will form a complex tone with its synchronous neighbour if the temporal to spectral ratio of the alternating tones are greater than the predetermined boundary (42.86 semitones/s).

The tones are of equal loudness so that only the spectral and temporal proximity are used for stream segregation.

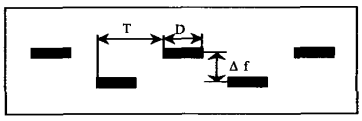

Fig. 1. Alternating Tone Series

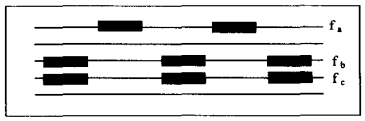

Fig. 2. Test Input

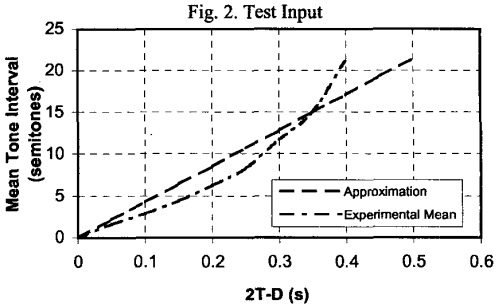

Fig. 3. Approximation to Van Noorden's Result

III. Calculating The TFD

Numerous techniques are available for calculating TFDs. These include FFT, DCT, MDCT, gamma-tone filterbank and other filterbanks. In much of the work in CASA the gammatone filterbank is used to extract a TFD. There is now a descriptor available from MPEG-7, the Audio Spectrum Envelope, which calculates a logarithmically spaced TFD, with similar spread of centre frequencies to the gamma-tone filterbank (See Fig. 4). The computational complexity of extraction is far less than the gamma-tone filterbank, as the Audio Spectrum Envelope is based on an FFT. Three different TFDs are used as a basis for a fuzzy system for stream segregation, exploiting the temporal coherence boundary. Each of the chosen three TFDs are explained below. The centre frequencies of the bands of the three different TFDs are shown in Fig. 4. As illustrated the MPEG-7 and the Gammatone curves are quite similar, although at low frequencies the spectral resolutionof the Gammatone filterbank is noticeably greater. Fig. 4, also shows that using the same number of bands in the linear method as the other methods gives poor low frequency resolution.

Each distribution used the same temporal resolution. A 10ms interval was used, the default value in the MPEG-7 Standard.

Gamma-tone filterbank

The gamma-tone filterbank represents the frequency response of the human ear, which is a consequence of the motion of the basilar membrane [15]. Many of the CASA systems proposed use the gamma-tone filterbank for pre-processing of raw auditory data [3][5][13]. The gamma-tone filter bank uses equivalent rectangular bandwidths (ERB). The implementation chosen cascades four second-order filters for each channel [16].

FFT – Linear Spectrogram

Although research has determined that the frequency response of the auditory system is best represented by the gamma-tone filter and ERB spaced channels (approximately a logarithmic response) there

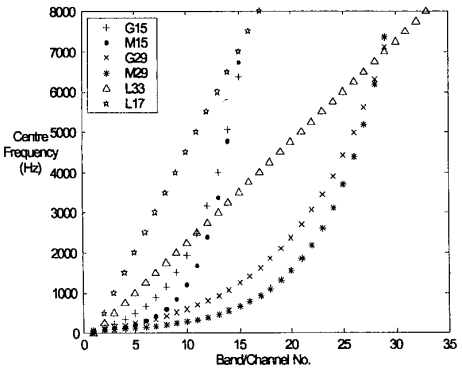

Fig. 4. Channel Centre Frequencies for Different Front-ends at selected resolutions. Including the MPEG-7 Audio Spectrum Envelope using '1/2 octave' (M15) and '1/4 octave' (M29) resolution, the Gammatone filterbank centre frequencies for 29 (G29) and 15 (G15) channels and linear distributed spectral frequencies for 33 (L33) and 17 coefficients (L17)

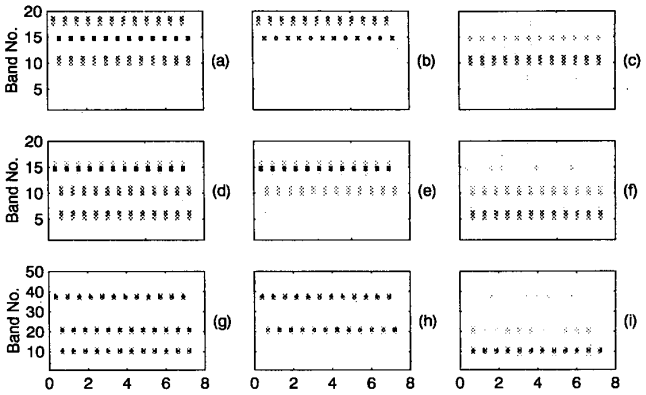

Fig. 5. Fuzzy System results, dark sections indicate sections of high spectral energy. (a) Original MPEG-7 Audio Spectrum Envelope TFD, (b-c) clustered data. (d) Original Gammatone TFD, (e-f) clustered data. (g) Original linear TFD, (h-i) clustered data.

are advantages in using a linearly spaced spectrum. One specific advantage is in the recognition of harmonically related elements, where integer multiples of the fundamental frequency are perceptually grouped; this is particularly difficult to see on a logarithmic scale, but simple to extract from a linear representation. Another advantage of using a linear spectrogram is the simplistic technique for reconstruction.

However, the logarithmic boundary between temporal and spectral proximity still remains, and the problem posed is how to implement the boundary condition on a linear scale. The simple solution is to implement the logarithm in the scaling operation. To calculate a linear spectrum it is sufficient to use a sliding window FFT approach, a 30ms analysis window was used with 2/3 overlap, delivering 10ms temporal resolution.

MPEG-7 Audio Spectrum Envelope

The MPEG-7 Audio Spectrum Envelope is one of the low level descriptors (LLDs) defined in MPEG-7 Audio. The recommended implementation of this transform is to use a windowed FFT, using a 30ms analysis window and a 10ms shift, to deliver a linear spectrum and then the squares of the magnitudes of the coefficients (the power) in logarithmically spaced bins are summed to give a logarithmically spaced spectrum.

The versatility of this descriptor and transformation technique provides extensive opportunities for practical application. The spectral resolution may be specified from $1/16^{th}$ of an octave to 8 octaves. By using various resolutions different information may be extracted, from extremely detailed accounts to a simple within band versus out-of-band power. One other advantage of the MPEG-7 Audio Spectrum Envelope is the compactness of the spectral description, especially compared to the linear spectrogram with similar low-frequency resolution. In a two second section with a '1/4 octave' resolution, assuming a 16kHz sampling frequency, will give 28 MPEG-7 Audio channels, with 2 additional out-of-band channels. To get similar resolution at low frequencies in a linear representation requires around 240 linear spectral coefficients, requiring 8 times the amount of data storage. Over large quantities of data and lengthy audio signals the information storage becomes difficult, however is far easier with the MPEG-7 Audio description.

IV. Results and Discussion

Generating Feature Vectors and the Threshold of Hearing

Given the amplitude of the tones (or at least the loudness) of the test data are all the same the amplitude variation is not required for clustering. The threshold of hearing may be employed to determine which elements of the TFD should be included as feature vectors for clustering. Since the test signals used had equal loudness a simple threshold decision was employed to determine inclusion of TFD elements in the feature vector set. If the amplitude of the TFD element exceeded the nominal threshold, then the time and frequency information of that element created a feature vector, otherwise the element was ignored.

Clustering Performance

Each different TFD created suitable feature vector sets that enabled a fuzzy clustering system to perform successful streaming of the test input, based on the temporal coherence boundary.

This is noticed particularly in the situation where informal listening tests indicate partial allocation of the central tone to each stream (see Fig. 5). The auditory system identifies that both groupings are possible. Each fuzzy system successfully performs a partial allocation of these TFD elements to each cluster. Fig. 5 (b-c), illustrate the remarkable performance of the clustering system using the MPEG-7 Audio Spectrum Envelope distribution, which clearly performs better than the other TFDs on the borderline.

Frequency and Time Scaling

The required scaling factors for the MPEG-7 Audio Spectrum Envelope and the Gamma-tone filterbank TFDs, with the same number of frequency channels, were assumed the same. The assumption was based on Fig. 4, which illustrated the similarity in the spectral spread of centre frequencies for the two TFDs. The scaling factor varied depending upon the number of channels used, based upon a semitone approximation to the channel bandwidths.

The performance of the clustering system indicates that appropriate scaling factors were used, with all three TFDs clustering sufficiently well. The frequency scaling of the linear TFD was completely different to the other TFDs. The log of the linear

frequency was used to convert to semitones and then the scaling factor could be applied.

With the MPEG-7 Audio Spectrum Envelope a '1/4 octave' resolution was chosen. Since each band covers a quarter of an octave and there are 12 semitones in an octave, each logarithmic band covers a 3 semitone interval. The TFD is calculated at 10ms intervals. So scaling is performed by multiplying the time index by 10ms and by multiplying the frequency values by 42.86/3

Each of the systems performed well with the approximated scaling, the MPEG-7 Audio Spectrum Envelope performed best. The poorer performance of the Gammatone filterbank may be attributable to the approximation of the channel frequency distributions; perhaps the 3-semitone approximation was not completely appropriate for the Gammatone filterbank. An alternate approach could use the centre frequencies of the filter channels to derive a semi-tone equivalent, similar to the linear approach.

The systems were also tested with other sets of scaling factors and the performance of the systems was poor in these test situations. Hence the psychoacoustic knowledge has proven true and useful in this situation. Psychoacoustic information is available for a number of other perceptual cues and should be able to be integrated into the fuzzy system in the future.

Execution Time and Meta-Data

Although the performance of the MPEG-7 Audio Spectrum Envelope TFD appeared best in clustering performance other factors need to be considered in the performance. The time required to cluster data from a 40s file varied considerably between the TFDs. For the gammatone filterbank the time was 125s, 11s for the linear TFD and 10s for the MPEG-7 Audio Spectrum Envelope. It is clearly apparent that the benefit of using the gamma-tone filterbank (as a good representation of the basilar membrane) is outweighed by the unfeasible execution time required, ensuring that computation could not be achieved in real time.

The MPEG-7 Audio Spectrum Envelope had an execution time 12 times faster than the Gamma-tone filterbank, substantially shorter than the length of the file and implies real-time analysis is possible.

The post processing to logarithmic values delayed the linear spectrum technique. A further problem with the linear TFD is the excessive amounts of meta-data. To store the clustered data approximately 8 times the amount of data storage is required, substantial proportions of this data is redundant.

For similar frequency resolution the number of TFD elements in the MPEG-7 Audio Spectrum Envelope and the Gammatone filterbank are identical, implying similar data storage requirements as meta-data. Overall the MPEG-7 Audio Spectrum Envelope outperforms the other two TFDs considered and correlates sufficiently with the human auditory system to be used in further development of a fuzzy system for audio stream segregation.

V. Conclusion

The time-frequency distributions examined in this work perform well in a fuzzy cluster system to examine stream segregation based on the temporal coherence boundary. By selecting appropriate scaling in each dimension the temporal coherence boundary is well defined. The selection of TFD depends primarily upon the versatility for further perceptual cues and the computational efficiency. Further work will aim to include other cues such as common amplitude modulation, amplitude proximity, common onset and offset, into the system to make a more robust stream segregation system. The MPEG-7 Audio Spectrum Envelope performs well for temporal and spectral proximity cues, however it provides a difficult platform for harmonic cues. The linear spectrum gives excessive information at high frequency, though it will allow easy implementation of harmonic cues. It should be noted that other MPEG-7 low-level descriptors may be able to accommodate harmonic cues, such as the Audio Fundamental Frequency or the Harmonic Spectral Descriptor, so the MPEG-7 Audio Spectrum Envelope is the conclusively best performer for stream segregation based solely on the temporal coherence boundary.

The ability to perform CASA with the MPEG-7 Audio Description Scheme would be a substantial advantage. CASA should ideally allow automatic transcription of audio files to high-level descriptions as defined in the MPEG-7 Audio description scheme. Hence, advancing the use of MPEG-7 descriptors is an advantage for the compatibility with this new international standard.

References

[1] R.H. Lambert, A.J. Bell, 'Blind separation of multiple speakers in a multipath environment,' *ICASSP-97*, pp 423-426, Vol. 1, 1997.

[2] U. Lindgren, A. van der Veen, 'Source separation based on second order statistics – an algebraic approach,' *8th IEEE Signal Processing Workshop on Statistical Signal & Array Processing*, pp 324-327, 24-26 June 1996.

[3] D.P.W. Ellis, *Prediction-driven computational auditory scene analysis*. PhD thesis, MIT, Cambridge, Massachusetts, USA, June 1996.

[4] D. Godsmark, G. J. Brown, 'A blackboard architecture for computational auditory scene analysis,' *Speech Communication*, Vol. 27, pp 351-366, 1999.

[5] F. Klassner, V. Lesser, S.H. Nawab, 'The IPUS blackboard architecture as a framework for computational auditory scene analysis,' *IJCAI-95 Workshop on Computational Auditory Scene Analysis*, Montreal, Canada, August 1995.

[6] Bregman, A.S., *Auditory Scene Analysis*, MIT Press, Cambridge, MA, 1990.

[7] Van Noorden, L.P.A.S., *Temporal Coherence in the Perception of Tone Sequences*, PhD Dissertation, Eindhoven: Technische Hogeschool Eindhoven, 1975.

[8] Moore, Brian C.J., *An introduction to the psychology of hearing*, 4th ed., London; San Diego, Calif. Academic Press, 1997.

[9] M. Slaney, R.F. Lyon, 'On the importance of time – a temporal representation of sound,' In: M. Cooke, S. Beet, M. Crawford (Eds.), Visual Representations of Speech Signals, pp 95-116, John Wiley & Sons Ltd, 1993.

[10] Van Noorden, L.P.A.S., 'Rhythmic fission as a function of tone rate,' IPO Annual Progress Report 6, 1991.

[11] Van Noorden, L.P.A.S., 'Minimum differences of level and frequency for perceptual fission of tone sequences ABAB,' *J. Acoustical Society of America*, Vol. 61, No. 4, April 1997.

[12] Neef, D.L., Jesteadt, W., Brown, E.L., 'The relation between gap discrimination and auditory stream segregation,' Perception and Psychophysics, pp 493-501, Vol 31. No. 5, 1982.

[13] Lehn, K.H., 'Modeling binaural auditory scene analysis by a temporal fuzzy cluster analysis approach', *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 1997.

[14] Bregman, A.S., and Ahad, P.A., *Demonstrations of auditory scene analysis*, MIT Press, 1996, Compact Disk.

[15] G. Kubin, W.B. Kleijn, 'On speech coding in a perceptual domain,' *ICASSP-99*, pp 205-208, Vol. 1, 1999.

[16] M. Slaney, 'Auditory toolbox version 2', Technical Report #1998-010 Interval Research Corporation, 1998.