IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. V (May-Jun. 2014), PP 44-48
www.iosrjournals.org

Segmentation to Sound Conversion

[1]Anindita Chatterjee, [2]Himadri Nath Moulick
[1]*Tata Consultancy Services, Kolkata, West Bengal*
[2]*Aryabhatta Institute of Engineering, Durgapur*

Abstract: *Our motive, the task of unsupervised topic segmentation of speech data operating over raw acoustic information. In contrast to existing algorithms for topic segmentation of speech, our approach does not require input transcripts. Our method predicts topic changes by analyzing the distribution of reoccurring acoustic patterns in the speech signal corresponding to a single speaker. The algorithm robustly handles noise inherent in acoustic matching by intelligently aggregating information about the similarity profile from multiple local comparisons. Our experiments show that audio-based segmentation compares favorably with transcript based segmentation computed over noisy transcripts. These results demonstrate the desirability of our method for applications where a speech recognizer is not available, or its output has a high word error rate. Also, this paper describes methods for automatically locating points of significant change in music or audio by analyzing local self-similarity. This method can find individual note boundaries or even natural segment boundaries such as verse/c hours or speech/music transitions, even in the absence of cues such as silence. This approach uses the signal to model itself, and thus does not rely on particular acoustic cues nor requires training. Here present a wide variety of applications, including indexing, segmenting, and beat tracking of music and audio. The method works well on a wide variety of audio sources.*

Keywords: *Segmentation, sound, acoustic pattern, speech recognizer, signal*

I. Introduction

An important practical application of topic segmentation is the analysis of the spoken data. Paragraph breaks, section markets and other structural cues common in written documents are entirely missing in spoken data. Insertion of these structural markers can benefit multiple speech processing applications, including audio browsing, retrieval, and summarization.

A variety of methods for segmentation have been developed in 1999 by Beefereman, 2003 by Dielmann and 2005 by Renals. These methods typically assume that a segmentation algorithm has access not only to acoustic input but also to its transcript. It is natural for applications where the transcript has to be computed as part of the system output, or it is readily available from other system components. However, for some domains and languages, the transcripts may not be available, or the recognition performance may not be adequate to achieve reliable segmentation.

In this paper, we explore a method for topic segmentation that operates directly on a raw acoustic speech signal, without using any input transcripts. This method predicts topic changes by analyzing the distribution of reoccurring acoustic patterns in the speech signal corresponding to a single speaker. In the same way that unsupervised segmentation algorithms predict boundaries based on changes in lexical distribution, our algorithm is driven by changes in the distribution of acoustic patterns. The central hypothesis here is that similar sounding acoustic sequences produced by the same speaker correspond to similar lexicographic sequences.

Analyzing high-level content structure based on low-level acoustic features poses interesting computational and linguistic challenges. For instance, we need to handle the noise inherent in matching based on acoustic similarity, because of possible variations in speaking rate or pronunciation. Moreover, in the absence of higher-level knowledge, information about word boundaries is not always discernible from the raw acoustic input. This causes problems because we have no obvious unit of comparison. Finally, noise inherent in the acoustic matching procedure complicates the detection of distributional changes in the comparison matrix.

We compare the performance of our method against traditional transcript-based segmentation algorithms. As expected, the performance of the latter depends on the accuracy of the input transcript. When a manual transcription is available, the gap between audio-based segmentation and transcript based segmentation is substantial. However, in a more realistic scenario when the transcripts are fraught with recognition errors, the two approaches exhibit similar performance. These results demonstrate that audio-based algorithms are an effective and efficient solution for applications where transcripts are unavailable or highly errorful.

II. Motivation

One of the first chapters of most textbooks in image processing or computer vision is devoted to edge detection and object segmentation. This is because it is much easier to build classification and analysis

algorithms using as input segmented objects rather than raw image data. In video analysis, shots, pans and generally temporal segments are detected and then analyzed for content. Similarly temporal segmentation can be used for audio and especially music analysis.

Auditory scene analysis is the process by which the human auditory system builds mental descriptions of complex auditory environments by analyzing mixtures of sounds [4]. From an ecological viewpoint, we try to associate events with sounds in order to understand our environment. The characteristics of sound sources tend to vary smoothly in time. Therefore abrupt changes usually indicate a new sound event. The decisions for sequential and simultaneous integration of sound are based on multiple cues. Although our method does not attempt to model the human auditory system, it does use significant changes of multiple features as segmentation boundaries. The experiments indicate that the features selected contain enough information to be useful for automatic segmentation.

Temporal segmentation is a more primitive process than classification since it does not try to interpret the data. Therefore, it can be more easily modeled using mathematical techniques. Being simpler it can work with arbitrary audio and does not pose specific constraints on its input like single speaker or isolated tones. It has been argued in that music analysis systems should be built for and tested on real music and be based on perceptual properties rather than music theory and note-level transcriptions.

Annotation of simple cases like musical instruments or music vs. speech can be performed automatically using current classification systems. Based on these techniques, a completely automatic annotation system for audio could be envisioned. Although not impossible in theory, there are two problems with such an approach. The first is that current systems are not perfect and, therefore, annotation errors are inevitable. This problem has to do with the current state of the art, so it is possible that in the future it will be solved. There is a second problem, however, that is more subtle and not so easy to address. Audio, and especially music, is heard and described differently by each listener. There are, however, attributes of audio that most listeners will agree upon, like the general structure of the piece, the style, etc. Ideally a system for annotation should automatically extract as much information as it can and then let the user edit and expand it.

This leads to a semi-automatic approach that combines both manual and fully automatic annotation into a flexible, practical user interface for audio manipulation. Automatic segmentation is an important part of such a system. For example, the user can automatically segment audio into regions then run automatic classification algorithms that suggest annotations for each region. Then the annotations can be edited and/or expanded by the user. This way, significant amounts of user time are saved without losing the flexibility of subjective annotation.

III. Speech and Music Segmentation

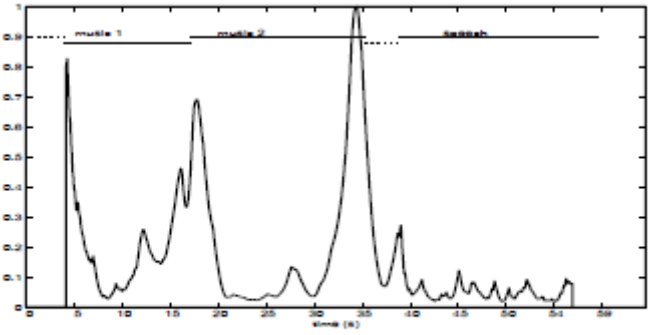
A variety of supervised and unsupervised methods have been employed to segment speech input. Some of these algorithms have been originally developed for processing written text in 1999 by Beeferman. Others are specially adapted for processing speech input by adding relevant acoustic features such as pause length and speaker change which is developed by Gallery in 2003, and by Dielmann & Renals in 2005. In parallel researchers respectively study the relationship between discourses structure and into national variation which was developed by Hirschberg and Nakatani in 1996 and by Shriberg in 2000. However all of the existing segmentation methods require as input a speech transcript of reasonable quality. In contrast, the method presented in this paper does not assume the availability of transcript, which prevents us from using segmentation algorithms developed for written text.

Also it is related to unsupervised approaches for text segmentation. The central assumption here is that sharp changes in lexical distribution signal the presence of topic boundaries in 2001 by Choi & in 1994 by Hearst. These approaches determine segment boundaries by identifying homogeneous regions within a similarity matrix that encodes pair wise similarity between textual units, such as sentence.

Beside music, these methods work for segmenting audio into speech and music regions. In the figure below, the audio novelty for the first minute of animals has young MPEG-7 content set. This segment contains 4 seconds of introductory silence, followed by a short musical segment with the production logo. At 17 seconds the titles start, and very different theme music commences. At 35 seconds, this fades into a short silence, followed by female speech over attenuated background music for the remainder of the segment. The figure shows the similarity score computed over a 8-second window, which is long enough to average out the short time spectral differences in speech. The largest peak occurs directly on the speech/music transition at 35seconds. The two other major peaks occur at the transitions between silence and music at 4 seconds and between the introduction and theme music at 17 seconds.

Segmentation by novelty score works well for musical notes and phases, as well as spoken phrases. Though difficult to qualitatively evaluate experiments using a variety of audio, such as TV dramas, yielded subjectively satisfying segmentation results. Even the soundtrack of a Warner Brothers "Looney Tunes" cartoon, containing a near-pathological sequence of orchestral sounds, sound effects, and Mel Blanc vocalizations was segmented with reasonable success. This segmentation method can't however be expected to segment speech into words

unless they are spectrally different. This is because words are often not well-delineated acoustically, for example the phrase "that's Steven" would be segmented into "that's-S" and "teven" because there is little acoustic differences in the "s" sounds of the two words, even with a glottal stop. Note that this would likely be the segmentation that a non-English speaker would choose.



Novelty score for video soundtrack

IV. Pattern Induction in Acoustic Data

It is a type of research on unsupervised lexical acquisition from continuous speech. These methods aim to infer vocabulary from unsegmented audio streams by analyzing regularities in pattern distribution developed by Marcken in 1996, by Brent in 1999 and by Venkataraman in 2001. Traditionally the speech signal is first converted into a string-like representation such as phonemes and syllables using a phonetic recognizer.

Park and Glass (2006) have recently shown the feasibility of an audio-based approach for word discovery. They induce the vocabulary from the audio stream directly, avoiding the need for phonetic transcription. Their method can accurately discover words which appear with high frequency in the audio stream. While the results obtained by Park and Glass inspire our approach, we cannot directly use their output as proxies for words in topic segmentation. Many of the content words occurring only a few times in the text are pruned away by this method.

V. Algorithm

The audio-based segmentation algorithm identifies topic boundaries by analyzing changes in the distribution of acoustic patterns. The analysis is performed in three steps. First, we identify recurring patterns in the audio stream and compute distortion between them. These acoustic patterns correspond to high-frequency words and phrases, but they only cover a fraction of the words that appear in the input.

As a result, the distributional profile obtained during this process is too sparse to deliver robust topic analysis.

Second, we generate an acoustic comparison matrix that aggregates information from multiple pattern matches. Additional matrix transformations during this step reduce the noise and irregularities inherent in acoustic matching.

Third, we partition the matrix to identify segments with a homogeneous distribution of acoustic patterns.

VI. Audio Segmentation and indexing

As demonstrated above, these methods give good estimates of audio segment boundaries. This is useful for applications where one might wish to access only a portion of an audio file. For example, in an audio editing tool, selection operations can be constrained to segment boundaries so that the selection region does not contain fractional notes or spoken phrases. This would be similar to "smart cut and paste" in a text editor that constrains selection regions to entire units such as words or sentences. The segment size can be adjusted to the degree of zoom so that the appropriate time resolution is available. When zoomed in, higher resolution would allow note-by-note selection while a zoomed out view would allow selection by phrase or section. Similarly segmenting audio greatly facilities audio browsing a "jump-to-next-segment" function allows audio to be browsed more rapidly than real-time. Because segments will be reasonably self-similar listening to a small portion will give a good idea of the entire segment. This would be especially appropriate when combined with a video shot detection system: shots having to have a significant audio novelty as well as video difference are more likely to be meaningful transitions. Another application might be to play back an audio piece synchronized with unpredictably timed events. Longer segments could be associated with particular stages, such as a game level or virtual environment location. As long the user stayed at that stage, the segment would be looped. Moving to a different stage would cause another segment to start paying.

VII. Audio Summarization and gisting

This approach can be extended to automatic audio summarization for example by playing only the start of each segment as in the "scan" feature on a CD player. In fact, segments can be clustered so that only significantly novel segments are included in the summary. Segments too similar to a segment already in the summary could be skipped without losing too much information. For example when summarizing a popular song repeated instances of the chorus could be excluded from the summary as they would be redundant. Reliably segmenting music by note or phrase allows substantial compression. For example, a repeated series of notes can be represented by the first note and the repetition times. The MPEG-4 structured audio standard supports exactly this kind of representation, but heretofore there have been few reliable methods to analyze the structure of existing of audio.

VIII. Audio feature selection

An important step of audio classification is feature selection. In order to obtain high accuracy for classification and segmentation, it is critical to select good features that can capture the temporal and spectral characteristics of audio signal. Based on the work, the features are divided into two types:

(i) Me frequency cepstral coefficients (MFCCs), and

(ii) Perceptual features. These features are combined as one feature vector after normalization.

Before feature extraction, an audio signal is converted into a general format, which is 8 KHz, 16-bit, mono-channel. Then it is pre-emphasized with parameter 0.98 to equalize the inherent spectral tilt and then divided into non-overlapping sub clips. A sub-clip is used as the classification unit, and segmentation is then performed based on the classification results. Classification performances with different durations of sub-clip are tested in our experiments. The sub-clip is further divided into non-overlapping 25 ms-long frames for feature extraction.

In our method, 8 order MFCCs are used as suggested by. The perceptual features we selected include: zero crossing rates (ZCR), short time energy (STE), sub-band powers distribution, brightness, bandwidth, spectrum flux (SF), band periodicity (BP), and noise frame ratio (NFR). While most of these features are frequently used, some are newly introduced, such as SF, BP and NFR. The definitions of these features are given below.

IX. Mel-frequency cepstral coefficients

These are computed from FFT. The log spectral coefficients are perceptually weighted by a non-linear map of the frequency scale, which is called Mel-scaling, using a triangular band pass filter bank. Then, the Mel-weighed spectrum is transformed into MFCC with the COS transformation.

$$cn = \frac{\sqrt{2}}{k \sum_{k=1}^{k} \left(\log[Sk) \cos\left[\frac{n(k-0.5)\pi}{K}\right] \quad n = 1,2, \dots \dots L\right]}$$

Where K is the number of band-pass filters, Sk is the Mel-weighed spectrum after passing kth triangular band-pass filter, and L is the order of the cestrum. In our method, 8-order MFCCs are used, that is L=8.

MFCC is commonly used in speech recognition system. Because of its good discriminating ability, it is also used in audio classification system.

Zero crossing rates

Zero-Crossing Rate is defined as the number of time-domain zero-crossings within a frame. It is simple measure of the frequency content of a signal:

$$ZCR = \frac{1}{2(N-1)} \sum_{m=1}^{N-1} |sgn[x(m+1)] - sgn[x(m)]|$$

Where sgn[.] is a sign function and x(m) is the discrete audio signal, m=1…N.

In general, speech signals are composed of alternating voiced sounds and unvoiced sounds in the syllable rate, while music signals do not have this kind of structure. Hence, for the speech signal, its variation of zero-crossing rate will be in general greater than that of music signals. ZCR is a good discriminator between speech and music. Considering this, many systems have used ZCR for audio classification.

X. Conclusion

Much prior work in audio segmentation has been based on detecting significant silence. Though this works satisfactorily for clean speech much common audio, such as popular music or reverberant sources, may contain no silence at all. The difference between a running average and a new spectral window is used to find "audio cuts" in though no results are presented to indicate how well this works. Another approach uses speaker identification to segment audio by speaker turns. Though the latter approach could be used to segment music, it relies on statistical models that must be trained from a corpus of labeled data, or estimated by clustering audio segments. We presented an unsupervised algorithm for audio based topic segmentation. In contrast to existing algorithms for speech segmentation, our approach does not require an input transcript. Thus, it can be used in domains where a speech recognizer is not available or its output is too noisy. Our approach approximates the distribution of cohesion ties by considering the distribution of acoustic patterns. Our experimental results demonstrate the utility of this approach: audio-based segmentation compares favorably with transcript-based segmentation computed over noisy transcripts.

The segmentation algorithm presented in this paper focuses on one source of linguistic information for discourse analysis — lexical cohesion. Multiple studies of discourse structure, however, have shown that prosodic cues are highly predictive of changes in topic structure (Hirschberg and Nakatani, 1996; Shriberg et al., 2000). In a supervised framework, we can further enhance audio-based segmentation by combining features derived from pattern analysis with prosodic information. We can also explore an unsupervised fusion of these two sources of information; for instance, we can induce informative prosodic cues by using distributional evidence.

References

[1]. H. M¨uller, N. Michoux, D. Bandon, and A. Geissbuhler,―A Review of Content-Based Image Retrieval Systemsin Medical Applications Clinical Benefits andFuture Directions,‖ International Journal of MedicalInformatics, vol. 73, pp. 1–23, 2004.

[2]. R.H. Choplin, J.M. Boehme, and C.D. Maynard,―Picture archiving and communication systems: anoverview,‖ RadioGraphics, vol. 12, pp. 127–129,1992.

[3]. Y. Liua, D. Zhang, G. Lu, and W.Y. Ma, ―A survey ofcontent-based image retrieval with high-level semantics,‖Pattern Recogntion, vol. 40, pp. 262–282, 2007.

[4]. [W. Hsu, S. Antani, L.R. Long, L. Neve, and G.R.Thoma, ―SPIRS: a Web-based Image Retrieval SystemFor Large Biomedical Databases,‖ InternationalJournal of Medical Informatics, vol. 78, pp. 13–24,2008.

[5]. H. M¨uller, T. Deselaers, E. Kim, C. Kalpathy, D.Jayashree, M. Thomas, P. Clough, and W. Hersh,―Overview of the ImageCLEFmed 2007 Medical Retrievaland Annotation Tasks,‖ 8th Workshop of theCross-Language Evaluation Forum (CLEF 2007),Proceedings of LNCS, vol. 5152, 2008.

[6]. T.M. Lehmann, B.B. Wein, J. Dahmen, J. Bredno, F.Vogelsang, andM. Kohnen, ―Content–based image retrieval in medical applications-A novel multi-step approach,‖Proc SPIE, vol. 3972, pp. 312-320, 2000.

[7]. R.B. Yates and B.R. Neto, Modern Information Retrieval,1st ed., AddisonWesley, 1999.

[8]. V. Vapnik, Statistical Learning Theory, New York,NY, Wiley; 1998.

[9]. T.F. Wu, C.J. Lin, and R.C. Weng, ―Probability Estimatesfor Multi-class Classification by Pairwise Coupling,‖Journal of Machine Learning Research, vol. 5,pp. 975–1005, 2004.

[10]. K. Fukunaga, Introduction to Statistical PatternRecognition, Boston, 2nd edition: Academic Press;1990.