Toward Content-Based Audio Indexing and Retrieval and a New Speaker Discrimination Technique

Lonce Wyse and Stephen W. Smoliar
Institute of Systems Science
National University of Singapore
Heng Mui Keng Terrace
Singapore 0511
(contact: lwyse@iss.nus.sg)

December 11, 1995

Abstract

Abstract: Several techniques for identifying segment transitions in an audio stream are discussed. Gross features are first identified that control more detailed and computationally expensive analysis down stream. Pitch is tracked using some basic streaming principles, and then used as one cue to speaker transitions. A novel speaker discrimination technique is described that makes segmentation decisions when a continuously updated model of the current speaker suddenly ceases to sufficiently account for the input data.

Keywords: Speaker discrimination, auditory streaming, audio stream segmentation, pitch tracking, nonlinear browsing.

Despite the multimedia hype, video and audio information are not currently part of our everyday computing environment. We don't yet have the tools for manipulating this kind of information with the ease with which we manipulate text. The goal of the Video Classification Project at ISS is to automatically segment a stream of image and sound data into meaningful units which can then be used in a database system [Smoliar and Zhang, 1994]. We consider the problems of parsing input streams, automatic indexing (labeling) of segments, and retrieval techniques. Such a system will support non-linear browsing of material and the use of sound and image keys for retrieval, which are far more natural ways of interacting with multimedia data than simple linear scanning. Currently, the audio and video stream parsing is done separately, however, the systems will run together, since each separate media stream contains information that can help the other make parsing decisions. The work reported upon here focuses on the parsing and indexing of the audio stream.

The immediate goal of the audio processing is to identify transition points between segments and to do an initial content oriented labeling of the segments. We use a combination of signal processing techniques for feature extraction, and "intelligent" symbolic level processing for decision making. The two processes work together in the sense that some hypotheses are formed based on initial signal processing work which in turn controls further signal processing to test, verify or refine the initial hypotheses.

The symbolic processing includes knowledge about characteristics of some of the basic signal types that we expect to encounter. Currently our data testbed is news stories, and the signals can

be grossly classified as "music", "speech" and "other" for the purpose of delineating meaningful segments. Speech is then further broken into segments at boundaries between different speakers.

1 Initial Processing

The signal is passed first through a filter which measures the amplitude envelope. Labels are attached to the signal identifying regions where energy is at some threshold percentage below the average where some of the following stages don't need to do any work. A 256 point FFT is then performed.

2 Music

The "music detector" is an extension of the work by Hawley [Hawley, 1993]. No deep philosophical issues about what music is are being addressed here. The system computes peaks in the magnitude spectrum, then bases its decision on the average length of time that peaks exist in a narrow frequency region. We improved upon previous work by using an ERB (Equivalent Rectangular Bandwidth) scaling of the frequency region [Moore and Glasberg, 1983]. Since this scaling is log-like above 500 Hz, it tends to be more robust than a linear representation because the sensitivity to peak movement is more uniform across frequency when the fundamental frequencies of speech or pitched musical instruments are non-stationary. Music detection is performed early because signals so labeled need no further analysis for our purposes. Sections that are not labeled as music are mined for more information.

3 Pitch

A spectrally-based pitch detection algorithm [Cohen, Grossberg, and Wyse, 1995] is employed which was designed to model aspects of human pitch perception. Also based on an ERB scaled energy representation of the signal, it employs an excitatory-center, inhibitory-surround mechanism that enhances peaks, and a weighted summation of regions around harmonics, to derive an activation strength function across pitch. It is robust under conditions of mistuned components, and models human responses to rippled noise and noise-band edges in addition to simple harmonic complexes.

The pitch model is layered, and includes a spectral representation, a contrast-enhanced spectral representation and finally a pitch layer, where, in general, every pitch has some level of activation. The pitch detector is robust against the effects of certain kinds of noise. Broadband noise is ignored, for example, even when the signal to noise ratio (in dB) is negative. Due to the convolution with the "Mexican hat" on-center off-surround kernel, spectrally broad signals are suppressed before influencing the pitch layer. More compact signals, particularly those with energy across several harmonically related components, are represented in the pitch layer, but unless they are specially constructed to do so, tend not to shift the peaks due to other pitched signals. This robustness to noise does not make this a model of multiple source segregation, however. Even a single tone creates many peaks in the pitch layer (at all subharmonics, though only one is maximal), so there is no obvious way to associate source perception with any but the most salient peak.

In order to track the pitch of voiced speech over time, several auditory streaming constraints [Bregman, 1990] are embedded in a following processing stage. Since the pitch detector responds to peaks and rippled noise, noise from fricatives cause peaks to appear in the pitch activation function, and, especially if the fricative is unvoiced, a noise peak can be the most prominent one. The resulting trace of the maximally activated pitch makes jumps that are too far in frequency and too fast in time for humans to track as a single stream. By incorporating constraints concerning the relationship between the distance and the rate of frequency jumps that result in a sequence of tones either streaming together or breaking into several streams, we are able to keep the pitch tracker following the pitch of just the voiced portion of speech. Similar constraints concerning energy keep the tracker from being distracted by low level pitched sounds or brief non-speech bursts.

4 Speech labeling

At this point in the pipeline, we have several representations of the signal and a stream of time-stamped labels. To label a segment as "speech," the next stage examines the pitch track in segments not already carrying a label incompatible with speech. The speech label begins with a pitched (assumed now to be "voiced") segment. The label ends with the last pitched segment before a time interval greater than one second in which no pitched segment lasted more than 75 ms. These criteria were empirically determined.

5 Speaker Discrimination

Speaker discrimination is an important component of segmenting an audio stream into meaningful subunits. Understanding when speakers change is crucial for dialogue understanding. In the realm of newscasts, a change in speaker almost always corresponds to a change in the the content, or news story. Speaker discrimination is related to speaker identification and verification, but the latter two processes are based on *a priori* knowledge about a limited number of speaker identities, and are usually text dependent. In speaker discrimination, only knowledge about speech in general is embedded in the system which is text independent. For the discrimination task, no matching of different segments (with speakers or with each other) is done, only temporally local decisions about speaker changes are made. Despite the fact that "inter-speaker variation" is the bane of speech recognition, actually extracting features that are invariant for one speaker, and that differ across speakers, is a challenging task.

Humans manage to recognize a change in speaker in a very short time, so averaging measures across tens of seconds should not be necessary. The methods used in our system combine pitch and spectral features, and make use of timing cues as well. Before the discrimination processes run, a segment must first be labeled as speech. Potential speaker transitions are flagged by events such as lengthy segments of non-speech, or sudden changes in pitch. Spectral features are extracted which are used for the final label assignment.

5.1 Pitch-based Speaker Discrimination

Changes in pitch characteristics make an important contribution to speaker discrimination, but are neither necessary nor sufficient for identifying the transition. The cue is perhaps most reliable when the transition is between speakers of different gender, but the overlap of ranges is

still considerable. Male vocal chords, tending to be longer and heavier than female's, generally produce fundamental frequencies in the range between 80-250 Hz, while those produced by females are generally in the range between 150-500 Hz. The range for children is slightly higher than that of for women.

Averaging of the speech signal over a window of time and looking for large changes in this measure is a possible technique, but we have found that pitch within a single utterance can vary widely even when averaged over a window of two or more seconds. Averaging also has the disadvantage of being too influenced by extremes; the more outlandish the greater the influence. We have therefore adopted the use of a change in pitch *range* for flagging possible speaker transitions.

The range has two frequency bounds: one above which a certain percentage of input pitch values lie, the other below which a certain percentage lie over the duration of a time window. If a cutoff percentage parameter is set at 50%, for example, the mean is tracked. We are currently using a cutoff of 25% for both the upper and lower range, and a window of 2 seconds. The actual frequency of outliers thus have no effect on the range computation no matter how outlandish, making this method more robust than averaging with particularly "prosodic" speakers.

The temporal localization ability of the range change discrimination technique is better than the window size, since it depends upon the cutoff percentages as well. With cutoff percentages of less than 50%, the high bound is more sensitive (responds more quickly) to an increase in upper range than to a decrease, and the low bound more sensitive to a decrease in the lower range than an increase. Changes to the range in the "sensitive direction" of the bound measures can happen as quickly as the cutoff percentage multiplied by the window length.

5.2 Spectrally-based Speaker Discrimination

Speaker variation is the bane of speech processors, and great pains are taken to normalize, compensate or otherwise make systems less sensitive them. Sources of variation include regional accents, emotional stress, speaking rate, physical impediments, health, gender, age, and chest, glottis and vocal tract morphology. With so much inter-speaker variability, it seems that automatic speaker discrimination should be easy. The difficulty, of course, lies in finding acoustic features that change less within a speaker than across speakers.

We are currently exploring a spectrally based method. It is even more true for spectra than for pitch, that no average over a window short enough for reasonably fast detection of speaker change, will be stationary over the course of a single speaker utterance.

One way to eliminate the effects of intra-speaker spectral variation would be to compare particular phonemes of one speaker to the same phonemes of another (recall that phonemes are linguistically, not acoustically, defined). This is one of the methods used in speaker identification and verification when comparing input to a known stored utterance [Furui, 1986]. There are several problems with this approach. First, it involves the identification of the phonemes. Since one of the manifestations of speaker differences is that different vowels spoken by the respective speakers can overlap in formant space [Peterson and Barney, 1952], then a fairly complete speech recognition system would have to be a part of the discrimination mechanism, and would thus carry a substantial computational burden. Another problem with this approach, is that speaker discrimination would be language dependent, and people can normally detect speaker changes even when a language unknown to the observer is being spoken.

Our method, related to this same-phoneme comparison method, is to break up a spectral space into regions, and compare new input only to stored data in the same spectral region that

the input belongs to. This eliminates the need for phoneme identification (though at the expense of being able to use that particular aspect of variation in the process), and turns the approach into a kind of spectral redundancy measure.

To break the representation space space into regions, we recorded 15 different voiced sounds (vowels, liquids and fricatives) at as close to a steady state as they could be spoken. Sixteen LPC-derived cepstral coefficients were taken using 25 ms windows stepped every 10 ms, and the vectors were averaged to produce one representative vector for each sound.

During the processing of segments of the input stream that have already been labeled by our system as both "speech" and "pitched", the most recent 750 ms (the "recency" window) of input vectors are compared to the previous 3 seconds (the "history" window), and a novelty score is computed. If the novelty score exceeds a certain threshold, then a new speaker is flagged as starting at the time corresponding to the beginning of the recency window.

The way the novelty score is computed is by first identifying the region into which each input vector falls, and then finding the closest vector in the history window already stored in that region. Euclidean distance, a standard for comparing cepstral vectors, is used. If this distance exceeds a threshold parameter, then the input vector is flagged as novel (see Fig. 1). If the number of vectors in the recency window that are flagged as novel is greater than a second threshold parameter (expressed as a percentage of the number of data points that the recency window holds), then the criteria for identifying a speaker transition is met.

A brief description of how the parameters were determined sheds some light on how the method works. The tessellation of cepstral space limits the range of history vectors that recent inputs are compared to. This is what prevents the intra-speaker spectral variation, due to different vowels, to influence the measure. Thus the number of regions must be large enough to prevent too many cross-vowel comparisons. Our tessellation corresponds roughly to the number of vowels and semivowels (glides and liquids) used in the English language. The number of regions must not be too large, lest there never be a "history" vector in the same region as the input for comparison. Similarly, the length of the history window, while needing to be as short as possible to achieve acceptable temporal resolution, needs to be long enough so that at any given time, there is a high probability that there are history window vectors in the same region as the input. The three seconds of "history" maintained provides reasonable assurance that much of the input will fall in regions with stored vectors. The recency window needs to be short for resolution, but long for robustness, and long enough so that history vectors in the same region are from a previous entry of the trajectory into the region. When the recency window is so short that this condition is not met, then the distance between a recency vector and a history vector is determined by their distance along the trajectory itself rather than a speaker-characteristic use of the region of cepstral space.

This spectral discrimination component of the system is still being developed, but preliminary results show some promise. Figure 2 (a,b,c) shows the running novelty score for three different speakers reading the same passage from a book. The maximum possible novelty score is 150 (if each of the 5 ms-spaced vectors in the .75 second recency windows was novel). The input for Figure 2 (d) was constructed from the first 12 seconds from the first reader, the second 12 seconds from the second reader and the final 12 seconds from the third reader. At each splice, the speaker changed in mid sentence (though not mid-word), while the natural flow of the text was maintained. The novelty scores following each speaker changes can be seen to reach a peak higher than any of the individual speaker scores.

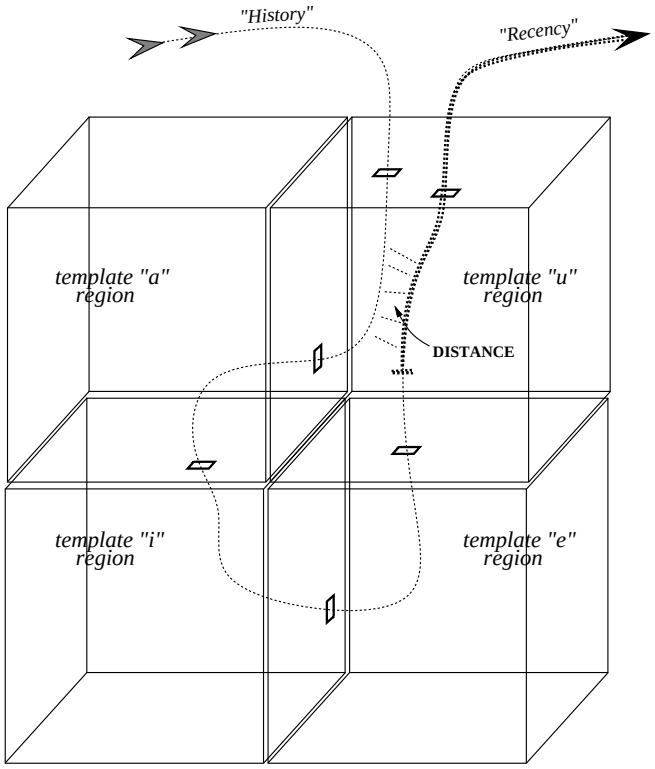

Figure 1: The successive cepstral vectors trace a trajectory in the tessellated space, and vectors within the "recency" window (most recent .75 seconds) are compared to those in the "history" window (extending back 3 seconds) that were in the same template region. If the distance between a recency vector and the closest history vector exceeds a threshold, it adds to the novelty score.

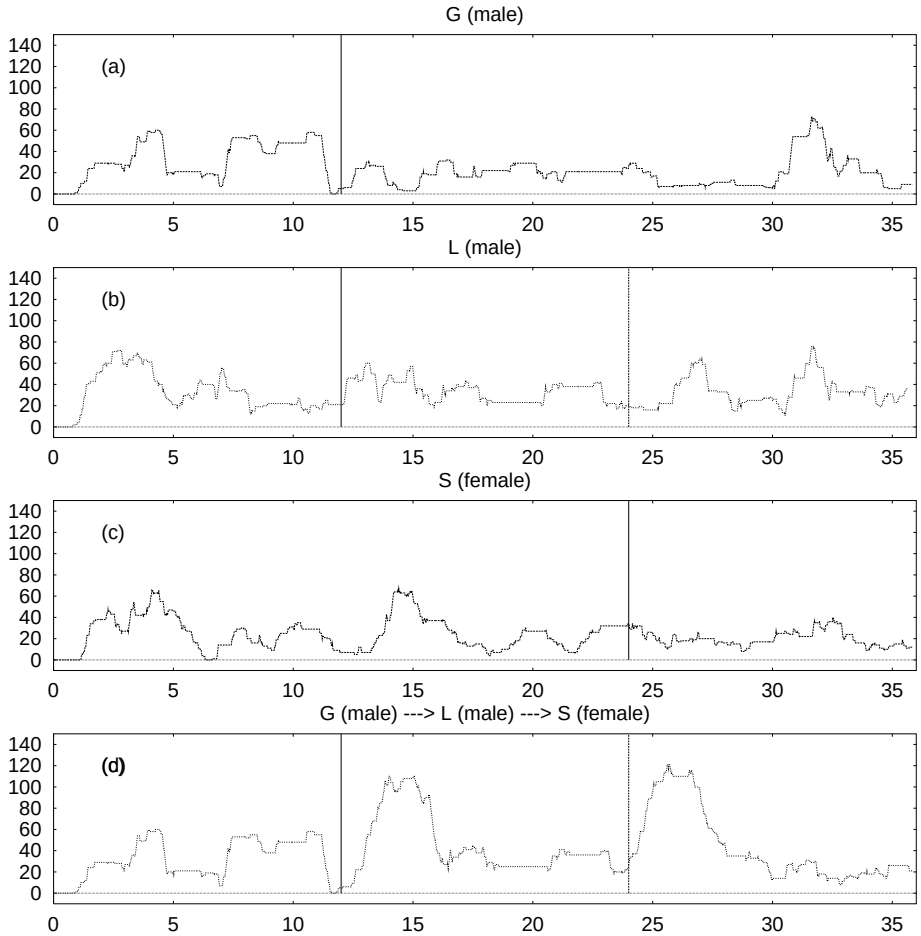

Figure 2: Novelty scores for 36 seconds of a paragraph read by (a) male speaker "G", (b) male speaker "L", and (c) female speaker "S". (d) The novelty score for the same paragraph spliced together from the first 12 seconds from "L", the second 12 seconds from "G", and the last 12 seconds from "S". The traces of the individual speakers can be recognized in (d) except just after the speaker changes where the score suddenly jumps.

Preliminary investigations suggest that the method described is fairly robust, although the variations within a single speaker can still produce novelty scores in the vicinity of speaker-change peaks. The technique has the advantage of having a relatively light computationally load since it requires no speech or phoneme recognition, and because the input vector is only compared to a fraction of the recent speech utterance. The computation other than the distance measures is minimal. The method might also be useful for text-independent speaker identification and verification by concatenating stored speech of a known speaker with a new speech signal and deriving the novelty score.

There are a number of ways in which this spectral method might be be improved. Using Perceptual Linear Predictive analysis rather than LPC derived cepstral coefficients may prove beneficial since distances between PLP vectors have been shown to correlate more consistently with perceptual distance [Hermansky, 1990]. The regions could be made adaptive with some continuously updated clustering method. While most likely improving the performance over the a priori and arbitrary division of the representation space, this would add considerable computation time. It also appears that some of our representation regions are proving to be more useful for discrimination than others (which is suggested by the observation that parts of the spectrum are more useful than others in identifying voice types [Bloothooft and Plomp, 1986]), and a systematic exploration of this will undoubtedly improve both the speed and the accuracy of this technique. Finally, other acoustic features that typically vary across speakers, but have a high intrapeaker variability with a high correlation to spectral variation (eg. spectral tilt, relative levels of even and odd harmonics, breathiness) might be more usefully compared by region in the manner described herein.

6 Discussion

In our video classification system, segment transition decisions in audio are based on less temporally localized information than are video transition decisions. However, all event labeling is done within 2 seconds of the event, and the processing runs in close to real time on a Sparc workstation, the actual run time being signal dependent.

The whole audio subsystem consists of some 20 or so different signal and symbol processing "filters" which can be run in a flexible ordering depending on the goals of the system. A uniform method of labeling and communication between processes has been developed which allows the information gleaned from one processes to control the processing parameters of another. Future work along these lines will make the flow of processing through the different filtering processes more flexible and integrated.

References

[Bloothooft and Plomp, 1986] Bloothooft, G. and Plomp, R. (1986). "Spectral analysis of sung vowels. III. Characteristics of singers and modes of singing," J. Acoust. Soc. Am. **79**, 852–864.

[Bregman, 1990] Bregman, A. (1990). *Auditory Scene Analysis* (M.I.T. Press, Cambridge).

[Cohen, Grossberg, and Wyse, 1995] Cohen, M., Grossberg, S., and Wyse, L. (1995). "A spectral network model of pitch perception," J. Acoust. Soc. Am. **98**, 862–879.

[Furui, 1986] Furui, S. (1986). "Research on individuality features in speech waves and automatic speaker recognition techniques," Speech Communication **5**, 183–197.

[Hawley, 1993] Hawley, M. J. (1993), "Structure out of sound," Ph.D. thesis, M.I.T.

[Hermansky, 1990] Hermansky, H. (1990). "Perceptual linear predictive (plp) analysis of speech," J. Acoust. Soc. Am. **87**, 1738–1752.

[Moore and Glasberg, 1983] Moore, B. and Glasberg, B. (1983). "Suggested formlae for calculating auditory filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753.

[Peterson and Barney, 1952] Peterson, G. E. and Barney, H. (1952). "Control methods used in a study of vowels," J. Acoust. Soc. Am. **24**, 175–184.

[Smoliar and Zhang, 1994] Smoliar, S. W. and Zhang, H. J. (1994). "Content-based video indexing and retrieval," IEEE MultiMedia **1**, 62–72.