Searching for Speech in Personal Audio

Daniel P.W. Ellis
LabROSA, Columbia University, New York

August 9, 2005

Abstract

Personal Audio, as might be collected by a continuously-recording flash-memory recorder worn on the body, presents an exciting frontier for search and access to details of an individual user’s daily activity. An MP3 player able to make such recordings – weighing a couple of ounces and able to record for over 10 hours on a single AA battery – can be bought today for under $100, but, while there is surely valuable information available in the complete soundtrack to one’s daily experiences, no tools currently exist to make such recordings remotely useful – since finding some event of interest entails review of minutes or hours of raw recordings. As part of a broader research project to investigate this modality and to develop search and retrieval mechanisms for it, we have identified detecting the absence or presence of speech as important attribute to take us towards a useful search mechanism for these data. This document proposes a project to develop an approach to identifying segments of speech in such recordings, which differ from the usual domain of speech detection (e.g. telephone calls) by the wide range of environments encountered (e.g. reverberation) and the very poor signal-to-noise ratios at which target speech would still be identified by a human listener (for instance, when engaged in a conversation in a busy restaurant). Our approach seeks to copy part of what listeners may be doing, by keying in to the pitch of a target listener and recognizing characteristically ‘speechy’ patterns of pitch and energy (prosodic) variation.

1 Background: Personal Audio

For the past two years we have been engaged in a project to develop mechanisms for accessing ‘personal audio’ – that is, recordings that attempt to capture everything a particular individual hears during daily life. Our initial focus has been on segmentation and classification at a coarse scale to identify the main ‘episodes’ that can form the basis of subsequent information search activities [Ellis and Lee, 2004a,b]: On a database of 62 hours, hand-marked into 139 segments each assigned to one of 16 broad classes, we were able segment the data based on changes in local properties, then cluster the resulting segments to achieve over 80% agreement with the ground truth at the level of the 1 sec feature frames. The result of this analysis is a diary-like representation, as illustrated in figure 1.

Part of the purpose of this project was to discover what uses such recordings might be put to, i.e. to find out what kinds of questions we would wish to ask or which items we would wish to retrieve were the data available. To an extent, our current ability to search in these recordings is still too crude to answer these
Figure 1: Screenshot from experimental browser/user interface. Recorded audio is shown by a pseudocolor spectrogram with a vertical time axis; clicking on the spectrogram begins playback from that moment, with fast-forward/reverse transport. In red next to the spectrograms are the automatically-derived segments along with their per-cluster manual labels. The green boxes show appointments read from the user’s online calendar, which prove a useful prompt in navigating the recordings.

questions, but the experience of collecting and processing more than one year’s worth of recordings (about 50GB) has given us insights into promising directions for our current research.

2 Finding Speech Segments

One very desirable feature would be the ability to identify all the segments in an audio archive containing a particular person’s voice. We had initially been looking at information other than voice in the recordings, thinking perhaps that something useful (like an automatic diary) could be created without having to provide access to the intelligible recordings themselves, with all the privacy issues that raises. However, it was clear from working with the recordings that voices – who they are and what they are saying – are the most interesting sound sources present in the recordings. Hence, we have decided to look into identifying these segments of speech (and to addressing the issues of privacy through a range of other approaches including scrambling of voices for which no consent-to-record is granted, and encryption of recordings).

We would like to be able to identify segments in which anyone is speaking, and where possible to identify who is speaking – a process called ‘diarization’ in the speech recognition community [National...
Institutes of Science and Technology, 2005]. Both of these functions – speech activity detection and speaker identification – are well established for telephony and broadcast audio, and have recently begun to be considered in the domain of meeting recordings [Pfau et al., 2001, Ellis and Liu, 2004], which have some resemblance to personal audio. However, existing techniques are completely inadequate to deal with the bulk of our data because of the very high levels of background noise and/or reverberation.

Figure 2 shows a typical example of the kind of noisy signal we would like to be able to handle. This is from a belt-mounted recorder worn during a discussion in a coffee shop. In this mono recording, several voices can be heard, but only for a word or two at a time – it is not possible to follow the conversation. It is, however, possible to identify the different speakers, given that they are familiar to the listener. (This example and several others can be heard from http://labrosa.ee.columbia.edu/speechsep/.)

Speech activity detection has been addressed in telephony (where detected inactivity can be exploited to reduce bandwidth) and speech recognition systems (since a recognizer will often find ‘words’ in crosstalk or background noise, leading to insertion errors). In the telephony domain the standard approach amounts to an energy threshold: there is no effort to distinguish between voice and other energetic signals. Speech recognition systems designed to work with broadcast audio must take a richer view and be prepared to exclude sounds such as music and other effects that may nonetheless have significant energy. The most successful approaches employ classifiers similar to, or based upon, the acoustic models of the speech recognizer itself to decide which segments resemble speech and are thus likely to be appropriate to pass on to the recognition engine [Williams and Ellis, 1999].

Neither of these approaches can be used for personal audio. There is no consistent energy level for the kind of speech we want to be able to detect, and the highly variable background noise will often be as loud or louder than target speech. And because of the significant noise background, the features used for conventional acoustic classifiers (e.g. Mel Cepstra) will represent a hopelessly entangled mixture of aspects of the speech and the background interference: short of training a classifier on examples of speech in every possible background noise we anticipate, any conventional classifier will have very poor performance.
3 Proposed Research

We are proposing a project to develop new techniques for identifying regions of speech in the kinds of high-noise, high-variability audio collected by body-worn continuous recorders. Our informal observation is that human listeners are particularly sensitive the dynamics of pitched, voiced segments as a way of detecting when voice is present, even in noisy backgrounds. This is consistent with our understanding of the speech signal, since the first few harmonics of the near-periodic vowels have the greatest energy of any part of a speech signal, and thus are the most likely to be detectible in poor signal-to-noise ratios (SNRs). Also, the redundancy of multiple harmonics derived from a single underlying periodicity gives rise to robust coding of the fundamental frequency for more accurate detection in noise.

Our approach is to adapt robust pitch tracking algorithms that have been specifically developed to mimic the ability of listeners to detect periodicity amidst noise [Ellis, 1997, Wu et al., 2003]. These algorithms perform a nonlinear combination of periodicity information in different spectral regions to best exploit locally-favorable SNRs, and to be able to identify periodicities present across the entire spectrum even when the evidence in any single frequency channel is weak.

This first stage will represent a noisy recording as a set of pitch fragments, with each fragment described by a pitch track and some kind of energy contour – either scalar, or with some crude spectral distribution. When the signal does indeed consist of noisy speech, each syllable in the speech is likely to result in a single fragment, with increasing continuity between syllables as the SNR improves. Detecting speech episodes then becomes the problem of looking at a collection of these fragments and deciding whether they are likely to have originated in speech, or from some other periodic source such as machinery, music, animal sounds etc. Psychological studies on intelligibility of highly distorted speech have suggested that there are some kind of ‘characteristic modulation patterns’ that typify speech [Remez et al., 1994]; while these patterns have not been defined, we believe that statistical classification based on the properties of the pitched fragments should be successful at detecting speech over windows of a few seconds (where shorter timespans contain too little information to support a decision). In particular, voiced speech consists of pitch tracks that vary neither too little (i.e. monotonous) nor too much (due to the limitations of the speech articulators). Both the pitch and energy prosodic contours within a phrase follow common general patterns, such as the “4 Hz” modulation peak resulting from the stream of syllables in fluent speech [Houtgast and Steeneken, 1973, Scheirer and Slaney, 1997].

4 Plan, Milestones, and Budget

The first stage of the research will be the development of the system to extract the pitched, syllable-like fragments from noisy recordings based on the perceptual pitch models. We will evaluate this algorithm on clean speech with noise artificially added, so that performance can be systematically compared to ground truth under controlled conditions. To verify that the approach also works for in our target domain, we will hand-mark speech syllables and their pitch tracks in some excerpts from our existing personal audio recordings, chosen to be the more challenging examples of speech we expect the approach to detect. Our
The next stage is to build the classifier that decides if a collection of pitched fragments contain evidence of speech, or if they arise from some other source. The essence of what we are trying to capture here is the broad parameters of pitch, energy, and timing variation observed in real speech but not in other environmental sounds. Thus, we can draw on the various large corpora of clean speech to train models of their expected properties; counterexample models can be extracted from our archive of personal audio recordings, automatically selected from regions that have been confirmed to contain no speech. The best features and statistics for this task remain to be identified, but we will investigate modulation spectra of smoothed pitch and energy tracks, as well as pruning schemes to identify voice-related fragments embedded in clutter from interfering noise. Our goal will be to achieve at least 90% correct in classifying individual segments as including speech or not, based on a test set drawn from the personal audio recordings but chosen to include both the more challenging examples of present-speech, and the most deceptive negative examples. This stage will also take three months and should result in a nice realization of the earlier hypothesis (from psychology) of the existence of specific low-level ‘speech patterns’ that are preferentially detected by human listeners. This work would make a good conference paper for the International Conference on Spoken Language Processing, to be held in Pittsburgh in September 2006.

The final stage of the project is to integrate the new detector within a system for search within complete long-duration recordings, both to tackle issues that arise from continuous, unsegmented recordings, and to exploit the new information within our prototype user interface. At this stage we will also begin to investigate the possibilities of identifying specific speakers through the clustering of different episodes of speaking based on higher-level statistics such as overall pitch distribution and speaking rate. Our primary goal here will be to achieve at least 70% recall at 90% precision for the speech segments which will be hand-marked in our original 62 hour test corpus. This stage will take a further 3 months and should result in a journal paper describing the overall project and how it fits in to our personal audio investigations.

The project will be conducted by one graduate research assistant working throughout the academic year, with one month of time budgeted for the Principal Investigator to provide supervision and other input. The total budget including stipend, tuition, travel to two conferences to present results, computational resources with which to conduct the project, and departmental overhead, amounts to $71,000.

5 Summary and Conclusion

Personal audio presents a fascinating opportunity of data that is easily and cheaply collected, that obviously holds a great deal of interesting information, but that lacks the adequate content analysis and search technologies to make it actually useful. Based on our preliminary experiments and data capture, we are proposing to focus specifically on the identification of spoken segments since these are generally the most
important and interesting for the user. However, the highly variable noise and channel conditions make this a considerable challenge that cannot be solved with conventional voice activity or speech detection techniques. Instead, we propose to mimic what appears to be a strategy used by human listeners, of focusing particularly on the pitched, voiced segments within speech, then deciding if a sequence of such fragments arise from speech or from some other source. There has been little or no prior work done on searching for and identifying speech in such high noise conditions, so this project will be breaking important new ground.

A successful approach to identifying regions of speech will add an important high-level attribute to our personal audio archives, allowing the system to automatically filter out the many long stretches without interesting content. Such a technology will also pave the way for the next stage of analysis, which is to cluster and identify recurring speakers in these recordings. We believe that the same prosodic features used to discriminate between speech and other sources will also form a solid basis for distinguishing between different speakers, a very useful attribute for searching within personal audio archives.

References


