Voice Activity Detection in Personal Audio Recordings Using Autocorrelogram Compensation

Keansub Lee and Daniel P. W. Ellis

LabROSA, Department of Electrical Engineering
Columbia University, New York, NY, 10027
{kslee,dpwe}@ee.columbia.edu

Abstract

This paper presents a novel method for identifying regions of speech in the kinds of energetic and highly-variable noise present in 'personal audio' collected by body-worn continuous recorders. Motivated by psychoacoustic evidence that pitch is crucial in the perception and organization of sound, we use a noise-robust pitch detection algorithm to locate speech-like regions. To avoid false alarms resulting from background noise with strong periodic components (such as air-conditioning), we add a new channel selection scheme to suppress frequency subbands where the autocorrelation is more stationary than encountered in voiced speech. Quantitative evaluation shows that these harmonic noises are effectively removed by this compensation technique in the domain of autocorrelogram, and that detection performance is significantly better than existing algorithms for detecting the presence of speech in real-world personal audio recordings.

Index Terms: voice activitity detection, pitch determination, noise robustness, environmental audio.

1. Introduction

Personal audio archives – continuous recordings of an individual's everyday experiences – can easily be captured by a flash-memory MP3 recorder worn on the body with low cost and high reliability [1]. While the collection of large personal audio archives provide a wide range of surely valuable information such as the daily locations and activities of the user, no tools currently exist to make such recordings remotely worthwhile – since finding a particular event of interest would require review of the entire raw recordings.

In our previous work we developed an automatic indexing mechanism at a large time-frame scale (e.g. 60 s) to identify the locations of a user based on the nonspeech background ambience statistics [2]. However, it has become clear that the richest and most informative content in these recordings is the speech, and thus it is important to be able to distinguish which segments of the sound contain speech via Voice Activity Detection (VAD). For example, dividing into speech and nonspeech allows both purer modeling of background ambience (for location recognition) and more focused processing of speech (for speaker identification, or for privacy protection by rendering detected speech unintelligible).

Most previous work on VAD has addressed the telephony domain, where standard approaches enhance a basic energy threshold; there is little 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. One approach is use a classifier based on the same representation used in the recognizer [3].

While these approaches are often quite effective in benign acoustical environments, e.g. a conference room, they tend to be less accurate in real-world, complex acoustic environments. Figure 1 (a) shows a typical example of personal audio we would like to be able to handle. There is no consistent energy level for the speech, and the highly variable background noise will often be as loud as or louder than target. Because of the significant noise background, features used for conventional acoustic classifiers (e.g. Mel Cepstra) represent a hopelessly entangled mixture of aspects of the speech and the background interference. As a consequence, unless we can train a classifier on examples of speech in every possible background noise we expect poor performance from any conventional classifier.

To detect regions of speech in this kind of high-noise, high-variability sound, we draw inspiration from the particular sensitivity of listeners to pitch, and to its dynamics. The first few harmonics of pseudoperiodic 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. As a result, our approach is based on a class of noise-robust Pitch Detection Algorithms (PDAs) that perform nonlinear combination of periodicity information in different spectral regions to best exploit locally-favorable SNRs, and can thus identify periodicity present across the entire spectrum even when the evidence in any single frequency channel is weak [4].

However, to use such PDAs to detect speech implicitly assumes that any periodicity present in the signal corresponds to voice. When the signal contains interference that is itself periodic – such as the steady hum of an air-conditioning unit, which is particularly common in some of our outdoor recordings – this approach to VAD raises many false alarms. In figure 1 (b), there are a fair number of obviously erroneous nonspeech pitches, as well as distortions of the voiced pitches, due to air-conditioning noise. Even multi-pitch trackers (like [5]) cannot separate such noise because voiced pitches are often weaker and/or intertwined (or overlapped) with non-voice, interfering pitch. Moreover, because these noises sometimes have higher spectral energy than speech, conventional spectral subtraction methods fail to estimate the correct local noise model for them and are thus unable to effectively eliminate them in the domain of spectral energy, as seen in figure 1 (c).

In the next section, we describe a new method to remove long-time stationary periodic noises in the domain of autocorrelogram seen in figure 1 (d). Based on the fact that the autocorrelation func-

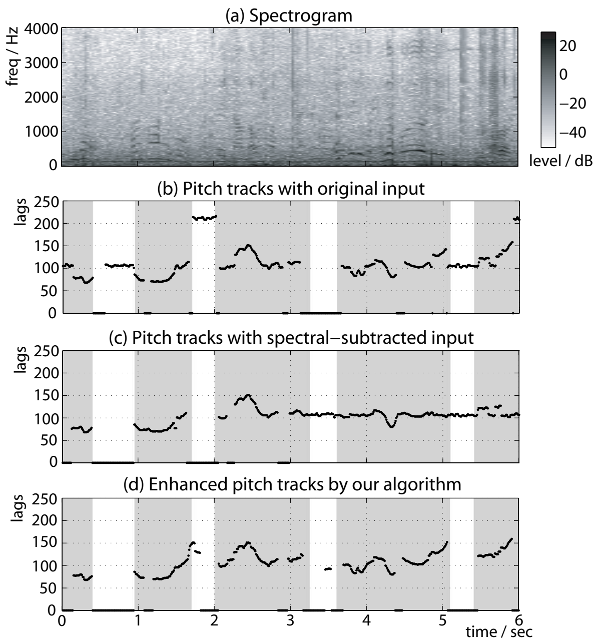

Figure 1: *Example of noisy speech from a personal audio recording. The pitch tracks in (b) and (c) are extracted by a noise-robust PDA as described in the text; pane (d) shows the result of our algorithm with the same input signal. The pitch of a stationary periodic air-conditioning noise appears as flat contours around lags 105 and 210 in (b), and tends to be more dominant around 4-6 s in (c) due to the failure of a noise estimation of the spectral subtraction, but is clearly deleted by our method in (d). Shadowed regions indicate manually-labeled voiced segments.*

tion (ACF) of these noises has a more slowly-changing shape compared to speech over long durations, subbands corrupted with such noise can be excluded from the summary autocorrelation (SAC) by estimating whether the current ACF and the local average ACF are similar. Evaluation and conclusions are presented in section 3 and 4 respectively.

2. Algorithm

Our system is based on a noise-robust PDA [5] that estimates dominant periodicities from an SAC formed by summing the normalized short-time ACFs of multiple subbands (based on a perceptual model filterbank, with 128 4^{th} order IIR gammatone filters uniformly spaced on the ERB scale). Critically, ACFs are excluded from the SAC if they appear to be dominated by aperiodic noise, so the SAC describes the periodicities present only in relatively noise-free portions of the spectrum, chosen frame by frame. Specifically, the SAC is built from only those subbands whose normalized ACF has a peak above 0.945, where a peak of 1.0 would corresponds to a perfectly periodic signal, and added noise reduces this value (this threshold was established empirically in [5]). Finally an HMM is used to extract the most probable pitch track from the SAC.

As described below, our modification is to further exclude channels in which similarity between the current ACF and its average over a longer time window exceeds a threshold automatically adapted to differentiate between dynamic periodic signals such as

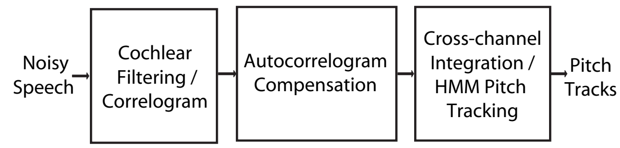

Figure 2: *Block diagram of our proposed system.*

voiced speech, and stationary periodic noises like air-conditioning. A simplified block diagram of our system is illustrated in figure 2.

2.1. Multichannel Autocorrelogram

Single-channel (mono) input recordings are resampled to 16 kHz, and then passed through a bank of gammatone filters uniformly spaced on an ERB scale. We used the channels spanning 80 Hz to 800 Hz to capture the strongest pitched-voice energy. Then, the envelope is calculated by half-wave rectifying these outputs.

The ACF $r_{yy}(c,n,\tau)$ and its energy $e_{yy}(c,n,\tau)$ for each subband envelope output $y(c,n)$ at a given frequency channel c and time index n may be defined as:

$$r_{yy}(c,n,\tau) = \sum_{i=n+1}^{n+W} y(c,i)y(c,i+\tau) \tag{1}$$

$$e_{yy}(c,n,\tau) = \sqrt{\sum_{i=n+1}^{n+W} y^2(c,i) \sum_{i=n+1}^{n+W} y^2(c,i+\tau)} \tag{2}$$

where W is an integration window size, and $r_{yy}(c,n,\tau)$ and $e_{yy}(c,n,\tau)$ are calculated over 25 ms windows every 10 ms for lag $\tau = 0 \ldots 400$ samples (i.e. up to 25 ms for a lowest pitch of 40 Hz). $r_{yy}(c,n,\tau)$ has a large value when $y(c,n)$ is similar to $y(c,n+\tau)$, i.e. if $y(c,n)$ has a period of P, then $r_{yy}(c,n,\tau)$ has peaks at $\tau = lP$ where l is an integer. The normalized ACF $r_{yy}(c,n,\tau)/e_{yy}(c,n,\tau)$ always falls between 0 and 1 (for our nonnegative envelopes), and thus a value of 1 at nonzero lag implies perfect repetition of a signal periodic within the window. To simplify notation, variables c, n, and τ are henceforth dropped.

2.2. Autocorrelogram Compensation

Let us assume that noisy speech y consists of a clean voiced signal s and stationary periodic noise n i.e. $y(c,n) = s(c,n) + n(c,n)$. In this case, the ACF given by:

$$r_{yy} = r_{ss} + 2r_{sn} + r_{nn} \tag{3}$$

For large W, if we assume that $n(c,n)$ is zero mean and uncorrelated with $s(c,n)$, so $r_{sn} = 0$ i.e. $r_{yy} = r_{ss} + r_{nn}$. Taking the expected value of both sides gives:

$$E\{r_{yy}\} = E\{r_{ss}\} + E\{r_{nn}\} \tag{4}$$

Given an estimate of the autocorrelation of the noise $\hat{r}_{nn}$, we could derive an estimate of the uncorrupt speech signal as:

$$\hat{r}_{ss} = r_{yy} - \hat{r}_{nn} \tag{5}$$

2.2.1. Linear compensation

Theoretically, the ACF of a stationary periodic noise r_{nn} could be estimated during periods when the speech is inactive and then subtracted (or cancelled) from the ACF of the current frame r_{yy} resulting in the ACF of the clean speech $\hat{r}_{ss}$. However, there is no simple way to detect pure-noise segments in a highly noisy signal. Instead, we introduce a new method based on our assumption, supported by observation, that r_{nn} for the kinds of noise we are trying to remove changes very little with time. Consequently, the long-time average of the ACF r_{yy} tends to be close to r_{nn}. Thus, we can attempt to estimate the autocorrelation of the less stationary voice signal by, for each time frame and each channel, estimating $\hat{r}_{nn}$ as the average ACF over M adjacent frames $avg\{r_{yy}\}$, and then subtracting it from r_{yy}:

$$\hat{r}_{ss} = max(0, r_{yy} - avg\{r_{yy}\}) \tag{6}$$

where $max()$ ensures that the estimated ACF cannot be negative.

Compared with the original SAC, the stationary periodic noise is effectively suppressed in a linear-compensated SAC, as shown in figure 3 (b), but at the cost of some speech information, particularly at lags below 100 samples. The basic assumption on this linear compensation is that the expected (average) value of r_{ss} in equation 4 is zero. However, since autocorrelations of bandlimited signals will always be positive in the vicinity of zero lag, r_{ss} does not have a zero-mean distribution, and $avg\{r_{yy}\}$ does not provide an unbiased estimate of r_{nn} for these lags. As a result, even with a large averaging window (e.g. 10 s), our estimate of the noise ACF is greater than the actual value of the distortion at these lags, and thus some speech information is removed by the compensation.

2.2.2. *Non-linear compensation*

To avoid the noise over-estimation problems of linear compensation, for each time frame and each channel, we compare every r_{yy} to $avg\{r_{yy}\}$ by cosine similarity, and use this to make a hard decision to include or exclude that ACF from the SAC. If the similarity is greater than a threshold θ_1, the subband is considered noisy for that frame, and is thus excluded from contributing to the SAC.

$$k = Sim_{cos}(r_{yy}, avg\{r_{yy}\}) \tag{7}$$

$$\hat{r}_{ss} = \begin{cases} r_{yy} & \text{if } k \le \theta_1 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $Sim_{cos}()$ is the cosine similarity (dot product divided by both magnitudes) between the two ACF vectors.

θ_1 is automatically tuned based on voice pitch dynamics and harmonic spacing. Changes in target pitch cause r_{ss} to be smoothly varying along time, making r_{yy} differ from $avg\{r_{yy}\}$. Channels containing clean speech will thus exhibit local-minima in similarity k compared to their noise-dominated neighbors. Since voiced speech spectra will have equidistant harmonics with noise energy in-between [6], during speech segments, we may see clean voiced ACFs with noisy ACFs between them. If speech is corrupted by stationary, periodic noise, ACFs dominated by this noise are likely to persist in some channels over long time frames. Therefore, θ_1 is chosen as the mean of a set of cosine similarity values of entire channels over M frames. Decreasing the value of M makes it easier to identify periodic noise with shorter duration (or some variability), but risks making gross errors of mistaking speech with small pitch variation as background noise. A value of $M = 100$ (e.g. 1 s window) is a good compromise between robustness and the ability to catch short-duration stationary harmonic noises.

After excluding the frequency bands judged to be dominated by periodic noise, the SAC is calculated based only on channels

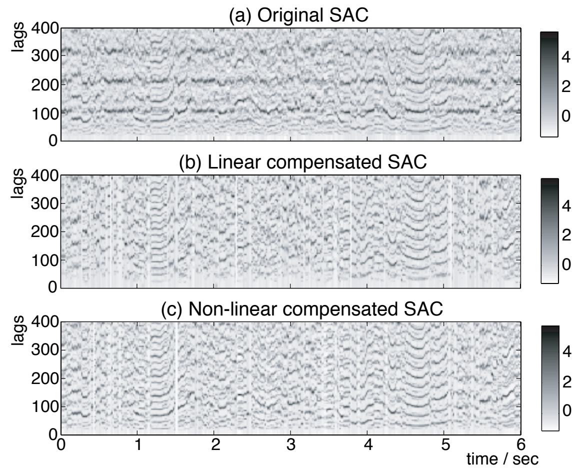

Figure 3: *SACs for the input signal from figure 1 with and without compensation using the local-average ACF over a 1 s window. Stationary harmonic air-conditioning noise appears as a sequence of strong peaks at lags of 105, 210 and 315 samples in the original SAC, but is clearly deleted in the non-linear compensated SAC (panel (c)), which also preserves speech information lost in the linear compensated SAC of panel (b). The non-linear compensated SAC is the basis of the enhanced pitch tracks shown in figure 1 (d).*

with a strong peak in the normalized ACF that exceeds a second threshold θ_2 (e.g. 0.945). θ_2 is chosen by examining the statistics from sample utterances mixed with interference [5]. Thus, the selected normalized ACF R_{yy} for every frame and channel is given by:

$$R_{yy} = \begin{cases} \hat{r}_{ss}/e_{yy} & \text{if } \hat{r}_{ss}/e_{yy} \ge \theta_2 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

2.3. Cross-channel Integration and HMM Pitch Tracking

As in [5], the R_{yy}s are integrated across frequency channels to obtain an SAC. Finally, an HMM is used to extract continuous pitch tracks. We define the pitch state as the union of two subspaces, one pitch or no pitch. In each frame, a hidden node represents the set of observed peaks. While the transition behavior with the same pitch subspace is modeled by a Laplacian distribution, the transition between different subspaces can be determined by training given a constant probability of a zero pitch. The Viterbi algorithm is used to find the most likely sequence of pitch states. We allow the probability of the no pitch state to vary according to the level of noise. Given a transition matrix estimated for relatively clean speech, we calculate pitch tracks with multiple different values for the zero-pitch probability, set as the n^{th} percentile of the SAC in each frame, and then determine the best percentile value by training. We also used the complete set of HMM posterior probabilities across all thresholds as a feature vector for SVM classification (below).

3. Evaluation

A 15 min test set was collected by a belt-mounted recorder worn during an outdoor discussion with four people (in front of the campus library), and thus was highly contaminated by noises including other people's voices and air-conditioning noise. We manually

Table 1: *Voice detection performance. The accuracy rate is the proportion of voiced frames correctly detected, and d′ (threshold-independent measure of class separation). The best value in each row is shown in bold. The best threshold for zero-pitch probability was estimated as the* 61^{st} *percentile of the SAC for the Binary Decision with Pitch Tracks system.*

	Guessing (Accuracy)	Binary Decision with Pitch Tracks (Accuracy, d')		SVM Classification with HMM Posterior (Accuracy, d')	
		Without Non-linear AC Compensation	With Non-linear AC Compensation	Without Non-linear AC Compensation	With Non-linear AC Compensation
FS/BS+NS	51.7%	73.8%, 1.66	**83.9**%, 1.99	75.9%, 1.73	83.7%, **2.05**
FS+BS/NS	68.0%	76.9%, 1.26	**81.0**%, **2.07**	74.2%, 1.60	80.2%, 2.00
BS/NS	66.2%	57.8%, 0.48	**75.7**%, **1.24**	59.3%, 0.63	71.9%, 1.17
FS/NS	61.8%	79.4%, 1.74	**88.0**%, **2.44**	76.5%, 1.96	85.8%, 2.36

annotated it into three categories: foreground speech (FS), background speech (BS) and nonspeech (NS). In our experiments, we compared four discrimination tasks: FS versus BS+NS, FS+BS versus NS, BS versus NS and FS versus NS.

The data set was divided into a 5 min training and a 10 min testing set. For our experiments, we computed the pitch track contour and the HMM posterior probabilities using every 5^{th} percentile of the SAC at each frame as the zero-pitch probability. We used these features as the basis for two voice detector systems: For the first system, after choosing the best fixed zero-pitch threshold on training set, we took the presence of a non-zero pitch track as indicating speech. The second system detected speech with a 2-way SVM classifier based on the 20-dimensional feature set of the HMM posterior probabilities across all zero-pitch probability settings.

As shown in figure 1, within speech regions labeled manually, there are many unvoiced segments between prominent syllables or words. Using pitch to detect the presence of voice cannot, of course, directly recognize these unpitched speech segments, but we smoothed the output of the pitch detector with a 1 s median filter to provide labels more directly comparable to the hand-labeled ground-truth.

The overall performance on the testing data is presented in table 1 in terms of the accuracy rate and d' (a threshold-independent measure, taken as the separation between two unit-variance Gaussian distributions that would exhibit the same level of performance). For comparison, we also used a baseline of guessing all frames as a single class. The accuracy and d' with the non-linear ACF compensation are significantly better than those without, which improves FS/BS+NS discrimination by about 10% absolute, and BS/NS discrimination by about 20%. Thus, the proposed algorithm is effective even for weak speech. The decision based on nonzero pitch track was simpler and by almost every measure (marginally) superior to the SVM classifier, and is thus preferred on the basis of its lower computational cost.

4. Conclusions

In this paper, we have proposed a robust pitch detection algorithm for identifying the presence of speech in the noisy, highly-variable personal audio collected by body-worn continuous recorders. In particular, we have introduced a new technique for estimating and suppressing stationary periodic noises such as air-conditioning machinery. The performance of our proposed algorithm is significantly better than existing pitch detection systems for the kinds of data we are addressing. Subsequent informal experiments have revealed that the sustained notes of background music can also be removed by this technique, which is a direction for further investigation e.g. for applications involving the recognition of broadcast speech: Detected voice pitch can be used for harmonic filtering to remove much of the nonspeech energy, to provide a drop-in replacement ASR feature. The multipitch tracker may also be helpful to suppress weak background voices after deleting strong stationary harmonic noises; this aspect is also currently under investigation.

5. Acknowledgements

This work is supported in part by the Microsoft Digital Memories (Memex) program, the National Science Foundation (NSF) under Grant No. IIS-0238301, and the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

6. References

[1] D. P. W. Ellis and K. Lee, "Minimal-impact audio-based personal archives," in *Proceedings of the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE)*, New York, NY, October 2004.

[2] D. P. W. Ellis and K. Lee, "Features for segmenting and classifying long-duration recordings of "personal" audio," in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing SAPA-04*, Jeju, Korea, October 2004.

[3] G. Williams and D. P. W. Ellis, "Speech/music discrimination based on posterior probability features," in *Proc. Eurospeech-99*, 1999.

[4] D. P. W. Ellis, "The weft: A representation for periodic sounds," in *Proc. IEEE Int. Conf. Acous., Speech, and Sig. Proc.*, 1997, pp. II–1307–1310.

[5] M. Wu, D.L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 229–241, 2003.

[6] C. Ris and S. Dupont, "Assessing local noise level estimation methods: Application to noise robust ASR," *Speech Communication*, vol. 34, no. 1–2, pp. 141–158, 2001.