Robust Voice Activity Detector for Real World Applications Using Harmonicity and Modulation frequency

Ekapol Chuangsuwanich and James Glass

MIT Computer Science and Artificial Intelligence Laboratory,
Cambridge, Massachusetts 02139, USA
{ekapolc, glass}@mit.edu

Abstract

The task of robustly detecting distant speech in low SNR environments for automatic speech recognition is examined using a two-stage approach based on two distinguishing features of speech, namely harmonicity and modulation frequency (MF). A modified metric for harmonicity is used as a gating function to a set of parallel classifiers that incorporate MFs computed on different frequency bands. Performance is evaluated on both the frame-level discriminative power and also the system level ASR results on a real-world robotic forklift task. Compared to other previously proposed features such as relative spectral entropy, and classification strategies involving MFs, the combined approach shows good generalization across different kinds of dynamic noise conditions, and obtains a significant improvement on the false alarm rate at low speech miss rate settings. The overall ASR results also improved significantly compared to the ESTI AMR-VAD2, while reducing the number of false alarms by a factor of two.

Index Terms: voice activity detection, modulation frequency, harmonicity, human-robot interaction.

1. Introduction

Voice Activity Detection (VAD) is the process of identifying segments of speech in a continuous audio stream. VAD is often the first stage of a speech processing application, and is used to both reduce computation by eliminating unnecessary transmission and processing of non-speech segments, as well as reduce potential mis-recognition errors in such segments. Since the process of making binary decisions about the presence or absence of speech is error prone, VAD is often avoided by requiring the user to initiate speech recording (e.g., using push-to-talk, or tap-and-talk mechanisms). In this paper, we consider the task of giving commands to an autonomous forklift [1]. It is crucial for the forklift to be able to continuously listen for possible commands, especially ones related to safety. In this scenario, VAD becomes an important consideration.

VAD has received considerable attention from the research community. In high quality recording conditions, energy-based methods perform well (e.g., [2]). In noisy conditions however, energy-based measures often produce a considerable number of false alarms. For this reason, a large variety of other features have been investigated for use in noisy environments (e.g., [3, 4]). Such techniques often require tuning parameters to a particular noise environment for them to be effective, and have difficulties dealing with non-stationary or instantaneous types of noises that are frequent in our task.

When it is difficult to know the nature of the noise a priori, features that measure fundamental attributes of speech are highly desirable. Two such attributes are harmonicity and temporal energy modulation. Tucker proposed a VAD based on detecting the harmonic structure of speech and showed good results even at very low SNR conditions [5]. However, since harmonicity is a basic property of any periodic signal, it is not useful by itself. Measures of temporal rhythm have been explored via modulation frequencies (MFs) which measure the temporal rate of change of energy across different frequency bands [6, 7]. Bach et al. used a purely MF-inspired set of features for discriminating between speech and non-speech which gave good generalization to noise types not included in training [8].

In this paper, we report on our experiments in a real world speech recognition system with VAD that incorporates both harmonicity and MF-based features. We explore a variety of configurations and perform experiments for the task of detecting shouted speech in an outdoor application involving interaction between humans and an autonomous forklift under dynamic outdoor noise conditions. By combining harmonicity with MF-based techniques we ultimately are able to improve the recognition compared to other standard VAD algorithms we investigated on real-world data.

The rest of this paper is organized as follows. Section 2 provides the reasoning behind the implementation of the VAD based on the harmonic structure of speech. Section 3 explains the MF extraction and the classification techniques applied. Section 4 describes the database and experimental results with some discussion. Finally, in Section 5 we provide some concluding remarks.

2. Harmonicity

In low SNR situations such as the one shown in Figure 1, non-sonorant portions of the speech signal are typically the first to become inaudible and masked by noise. In contrast, the harmonics associated with the main formants in vocalic regions often have the best local SNR conditions and are the most robust to additive noise. Thus, detecting the harmonic structure in speech sonorants has long been recognized as a noise-robust means to detect candidate speech regions, even in low SNR conditions. For example, Booersma proposed a periodicity measure that used normalized autocorrelation [9]. For a stationary time signal $x(t)$, the autocorrelation $r_x(\tau)$ as a function of the lag τ is defined as

$$r_x(\tau) = \frac{\int x(t)w(t)x(t+\tau)w(t+\tau)dt}{\int w(t)w(t+\tau)dt}$$

where $w(t)$ is the windowing function (in this case, a Hanning window). The division by the autocorrelation of the window compensates for the windowing effect of the autocorrelation.

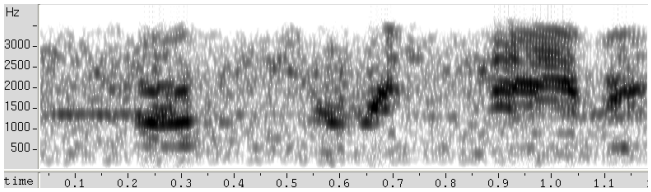

Figure 1: *Example of distant speech from the forklift database.*

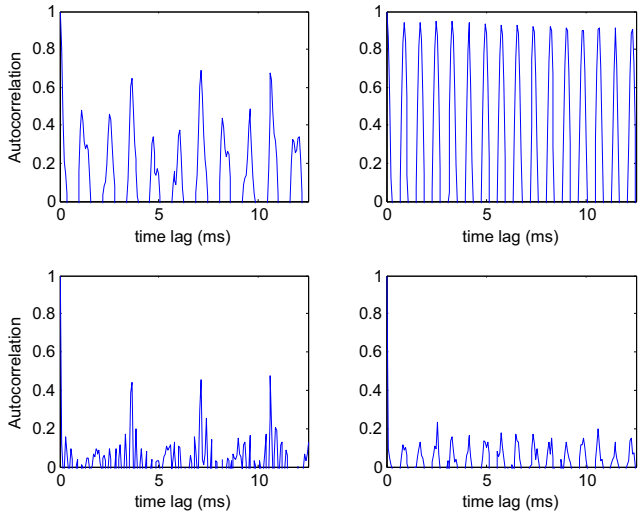

Figure 2: *Effect of cepstral bandpass liftering. Top left : Autocorrelation of a vowel with F_0 of 275 Hz. Top right : Autocorrelation of a 1200 Hz tone. Bottom left : Autocorrelation of the vowel after liftering. Bottom right : Autocorrelation of the tone after liftering.*

For a periodic signal, the highest local maxima will occur at the lag τ corresponding to the period. The relative power between the local peak and the zero-lag peak corresponds to the amount of periodicity in the signal. Harmonicity H is defined as:

$$H = 10 \log_{10} \frac{r_x(\tau)}{r_x(0) - r_x(\tau)}$$

However, since this periodicity measure will yield a high value for pure tones (e.g., such as beeping from a truck backing up), it is not practical in this basic form. In order to address this shortfall, bandpass cepstral liftering can be performed to extract frequencies corresponding to typical fundamental frequency values of human speech. Such filtering will retain the harmonic structure present in voiced speech, but filter out non-harmonic modulations such as spectral shape, tones, or periodic signals that are above human ranges (e.g., 1kHz). For our experiments, we extract frequencies in the 100-400 Hz range; however we observed that the results of our experiments were not overly sensitive to the precise values we used. Figure 2 shows the effect of the bandpass liftering on the autocorrelation function. Note how the pure tone now has a much lower local minima after the liftering.

3. Modulation Frequency

Harmonicity is able to capture local harmonic structure in the speech signal that is retained even in low SNR conditions. Figure 1 also shows that temporal energy fluctuations resulting from speech production are also present in the speech signal in such conditions. This basic modulation pattern around the 4Hz range is a fundamental property of speech [10]. One method to extract this information is via Modulation Frequencies (MFs) which extract frequency information over longer time spans in different frequency bands [6, 8, 7].

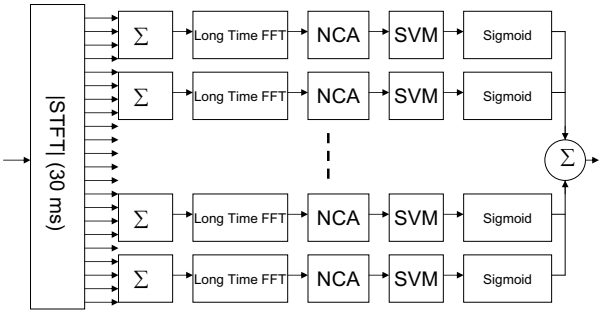

Figure 3: *The modulation frequency part of the VAD.*

The basic architecture of the MF-based VAD we explore in this work is illustrated in Figure 3. Essentially, it consists of an initial frequency analysis via Short-Time Fourier Transform (STFT) performed every 10ms with a Hamming window of 30ms. Individual STFT energy magnitudes are consolidated into a smaller number of N frequency bands by summing. The consolidated magnitude energies in each frequency band are then subjected to another STFT using a longer time window of M frames. A compression of $20 \log_{10}(1 + |x|)$ is applied to each MF to reduce the dynamic range. The resulting low-order STFT values correspond to the MFs in each frequency band. In our case, we typically use MFs in the range of 0 to 16Hz. Unlike [6, 8] who ignore the DC MF to remove convolutive noise, we found that the DC MF is an important cue for a speech event.

The output of the MF analysis can be used to perform a speech/non-speech decision every M frames. For this task researchers have explored the use of a single classifier [6, 8], as well as parallel classifiers that make independent decisions on each frequency channel [7]. Although Figure 3 shows the parallel SVM configuration, we explored and report on both methods in this paper. In our experiments we used a support vector machine (SVM) with radial basis kernels. Prior to SVM classification the low-order MFs in each channel are first transformed via Neighborhood Components Analysis (NCA) [11]. NCA is attractive since it does not make strong assumptions about the distribution of the underlying classes. We learn a single NCA transformation matrix and use it across all frequency channels.

The decision values from each SVM are passed through a compression function $1/(1+\exp(-2d))$ in order to avoid overconfidence in the decision of any particular frequency channel. The compressed decision values are then summed across all channels for a final decision value. This value is compared to a decision threshold, θ_{MF}, to decide the final classification label. The threshold can be used to trade-off missed detections versus false-alarms. In the case of our experiments we chose to operate with a low miss detection rate since we could pass the result to a speech recognizer to further process the result.

Although not shown in Figure 3, harmonicity was a necessary condition for a M frame segment to be classified as speech. The harmonicity value H was required to exceed a threshold, θ_H, for classification to proceed.

4. Experiments and Discussions

4.1. Speech/Non-speech detection capabilities

We performed a series of experiments to evaluate harmonicity and MF-based VAD on a challenging VAD task involving de-

System	EER(%)	FAR(%)
RSE [4]	7.92	18.21
LTSV [12]	20.10	71.75
GMM:MFCC	31.75	61.12
GMM:MF	7.4	19.5
Harmonicity	12.93	56.69
SVM:MF_{whole}	5.08	14.23
SVM:MF_{ch}	4.53	8.99
SVM:MF_{whole}+NCA	4.93	11.47
SVM:MF_{ch}+NCA	4.15	8.38
SVM:MF_{ch}+NCA+Harm	4.01	7.65

Table 1: *EER and FAR comparison between VAD configurations on the synthetic dataset.*

tecting distant speech for commanding an autonomous forklift. In this part of the experiment, we started with the investigation of the harmonicity and MF features in speech/non-speech classification as a pure detection task, i.e. to evaluate the systems' potential to differentiate between speech and non-speech. We collected a synthetic dataset consisting of speech commands from 26 subjects with added noise to simulate a variety of SNR values ranging from -5 to 15 dB. Both speech and noise were real data recorded with an array microphone. Noise data consisted of a variety of recordings we expected to encounter, including engine, street, loading dock, background talking, and environmental sounds such as wind, etc. Classification was performed on each frame without any additional post-processing steps. Frames that spanned a transition between speech and non-speech were excluded. We report the equal error rate (EER) between frame-level missed detections and false alarms, as well as the false alarm rate (FAR) that was obtained upon setting the miss detection rate at 1%. Training was done on 4 minutes worth of speech and 22 minutes worth of non-speech data excluded from the testing set.

For comparison purposes, we also include results on systems optimized for this task based on relative spectral entropy (RSE) [4], long-term signal variability (LTSV) [12], and statistical model-based VADs (such as one in [13]) using MFCCs and MF as features to Gaussian mixture models (GMMs).

Performance of a variety of VAD configurations on the synthetic data are summarized in Table 1. For each configuration, we show the achieved result after optimizing parameter settings (e.g., N, M etc) on the held-out data. In terms of the long window size, we examined a range of durations from 300-1300 ms and found that a window anywhere between 500 to 1000 ms gave reasonable results. Ultimately, we used a window size of 640 ms for all subsequent experiments. The GMM:MFCC had the lowest performance, while GMM:MF performs competitively. This might be due to the fact that MFCCs are very susceptible to noise. The RSE-based VAD performed better than the harmonicity system alone; However, the combination of MF and harmonicity based systems reduced the FAR by a maximum of 71% compared to the RSE system.

We then examined different MF configurations. The single SVM configuration (MF_{whole}) (based on [8]) performed worse than the parallel SVM configuration (MF_{ch}) which is more robust to corruption from band limited noises. We then examined the use of NCA transformations prior to the SVM models and found that NCA improved the performance in both cases. The best VAD performance were obtained when harmonicity was added as a gating function to the parallel SVM, reducing the FAR by 9% compared to the best system purely based on MF.

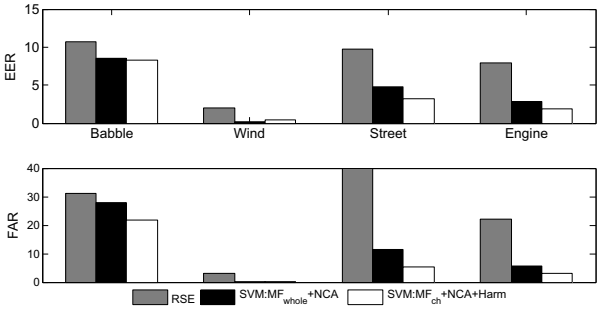

Figure 4: *EER (top) and FAR (bottom) performance for three VAD configurations on four noise conditions.*

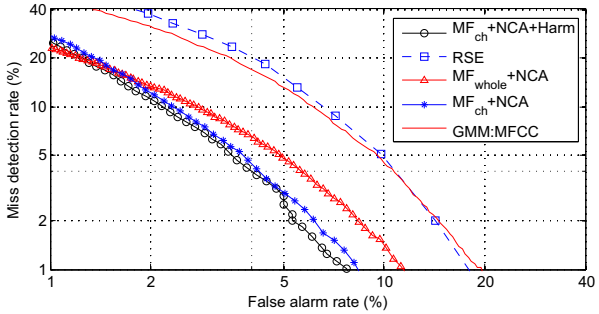

Figure 5: *Overall performance on the synthetic database.*

Since the synthetic dataset consisted of several different noise types, we also examined the performance of different configurations on specific noise conditions. The results are shown in Figure 4. From the figure we can see that the performance varied across specific kinds of noise, with babble noise being the most challenging. We also see that the parallel SVM+NCA configuration with harmonicity consistently did well in all noise conditions compared to the alternatives.

Finally, we observe that the EER and FAR values for each noise type are usually obtained at different thresholds. A system can do well on all noise types individually but poorly overall if the optimal thresholds for each noise type vary significantly. Table 1 and Figure 5 show the performance on all noise types with equal amount of testing utterances per noise type. Low EER and FAR values signify the potential of the system in terms of generalization to different environments without the need to re-tune. The harmonicity system has poor performance on its own since it is based on a simple metric. However, our measure is low in computational cost and works well as a pre-filter for the MF system. Sometimes the MF system on its own false triggers on a sequence of impulse-like sounds, such as hammer sounds, which can be removed by the harmonicity system.

4.2. Effect on ASR performance

In the second part of our experiments, we examined the influence of different VAD systems on the ASR results in the task of commanding an autonomous forklift in real world environments. Our evaluation data consists of actual recordings from four microphone arrays mounted on the forklift approximately 2.4 meters from the ground [14]. Recordings were made from 22 subjects commanding the forklift from distances typically tens of meters away in outdoor environments such as parking lots and warehouses. The SNR values ranged from 5 to 25 dB. The entire corpus, containing four microphone channels, is 10 hours long, with roughly 400 command words. The number of commands are sparse, since the forklift is mostly idle waiting for commands or taking the time to execute commands.

VAD	Words Correct (%)	WER (%)	FA (times/min)
Hand-labeled	56.3	45.0	0
RSE	44.1	59.6	0.70
AMR-VAD2	36.2	69.6	1.20
AFE	34.3	66.9	0.94
G.729B	7.0	93.5	0.30
MF+Harm	55.5	48.3	0.61

Table 2: *Word recognition results using different types of VAD.*

ASR was performed using PocketSUMMIT [15], a small footprint landmark-based speech recognizer. The task had a vocabulary size of 57 words. To filter out speech that was outside of the pre-determined set of commands, we also incorporated an explicit out of domain (OOD) command that was modeled by a single Gaussian mixture model trained on generic speech. OOD utterances in the database were tagged. The ASR model was trained on over 3,600 utterances of commands from 18 talkers under acoustic conditions similar to the corpus described in the previous paragraph.

Table 2 summarizes the ASR results. We showed the percentage of words correctly recognized and the word error rate (WER). The nature of the task requires continuous listening despite long periods of silence; as such the conventional definition of insertion error will yield unreasonable numbers. To have the WER correctly correlate with errors caused by the VAD clipping speech, triggering too early, or delaying too long (a "hang-over") after speaking has finished, we redefine insertion errors in this task to be those that occur within a one second vicinity of an underlying command. Utterances incorrectly recognized outside of the commands' vicinity are considered to be false alarms (FA). For example, if the reference command is "Stop," the hypothesis of "Stop it" in the vicinity counts as one insertion error. On the other hand, a hypothesis of "Stop it" in a period of silence counts as one false alarm at the utterance level instead of the word level. The number of false alarms is normalized by the length of the corpus, which represents the rate at which the forklift might improperly respond to false commands.

We compared our VAD with the other VAD systems described in the previous experiment with an inclusion of a simple hang-over scheme based on finite-state machines. All VADs were set to operate at 1% frame-level miss detection rate on the synthetic database. Standard VADs such as AMR-VAD2 [16], ESTI-AFE [17], and G.729B [18] were also considered. For reference, we also included hand-labeled boundaries from human experts as an ideal VAD to evaluate the ASR's effectiveness.

In terms of WER, the proposed system performed almost as well as an ASR system using hand-labeled end points. Although the accuracy was low, most of the recognition errors were style words which had little effect on the command level understanding. The proposed system also introduced the least amount of false alarms which was crucial for the robot to operate reliably. Note that the G.729B VAD remained on for extended periods of time, confusing the recognizer for OODs instead of commands. In the case of a recognition task in a noisy stream, the effect of a more accurate VAD outweighs the resolution lost by the frame size (in the MF system 320 ms). It is also possible to increase the resolution by doing a second pass with another type of VAD or voting with a smaller frame increment (such as one in [12]). This remains as an investigation for future work.

5. Conclusion

This paper described the task of VAD on distant speech in low SNR environments for an autonomous robotic forklift. Inspired by speech-like cues, namely harmonicity and MF, we designed a two-stage approach for speech/non-speech classification. Our experiments showed that the parallel SVM configuration usually outperformed classification based on the whole MF spectrum. A combination between MF and simple harmonicity measure helped reduce false alarm rate by another 9% at low miss rates. In ASR experiments, the proposed VAD outperformed standard VADs and achieved a WER very close to that of hand-labeled end points.

6. References

[1] S. Teller, M. Walter *et al.*, "A voice-commandable robotic forklift working alongside humans in minimally-prepared outdoor environments," in *Proc. IEEE Int. Conf. on Robotics and Automation*, 2010.

[2] C. Hsieh, T. Feng, and P. Huang, "Energy-based VAD with grey magnitude spectral subtraction," *Speech Communication*, vol. 51, no. 9, pp. 810–819, 2009.

[3] T. Kristjansson, S. Deligne, and P. Olsen., "Voicing features for robust speech detection," in *Proc. Eurospeech*, 2005.

[4] A. Ouzounov, "Robust features for speech detection - a comparative study," in *Int. Conf. on Computer Systems and Technologies*, 2005, pp. 19/1–19/6.

[5] R. Tucker, "Voice activity detection using a periodicity measure," in *Proc. IEEE*, vol. 139, 1992, pp. 377 –380.

[6] N. Mesgarani, M. Slaney, and S. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," in *IEEE trans. on Audio, Speech, and Language processing*, vol. 14, 2006, pp. 920–930.

[7] H. You and A. Alwa, "Temporal modulation processing of speech signals for noise robust ASR," in *Interspeech*, 2009, pp. 36–39.

[8] J. Bach, B. Kollmeier, and J. Anemuller, "Modulation-based detection of speech in real background noise: Generalization to novel background classes," in *ICASSP*, 2010, pp. 41–44.

[9] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. the Institute of Phonetic Sciences*, 1993, pp. 97–110.

[10] R. Drullman, J. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *JASA*, vol. 95, pp. 1053–1064, 1994.

[11] J. Goldberger, S. Roweis *et al.*, "Neighbourhood components analysis," *Advances in Neural Information Processing*, vol. 17, pp. 513 –520, 2005.

[12] P. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice acitivity detection using long-term signal variability," *IEE Trans. Audio, Speech and Language Processing*, vol. 19, pp. 600–613, 2011.

[13] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, 1999.

[14] E. Chuangsuwanich, S. Cyphers *et al.*, "Spoken command of large mobile robots in outdoor environments," in *SLT*, 2010.

[15] I. Hetherington, "PocketSUMMIT: Small-footprint continuous speech recognition," in *Proc. Interspeech*, 2007.

[16] ETSI, "Voice activity detector (VAD) for adaptive multi-rate (AMR) speech traffic channels," *ETSI EN 301 708*, 1999.

[17] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms dsr advanced front end," *ETSI ES 202 050*, 2007.

[18] A. Benyassine, E. Shlomot *et al.*, "ITU-T recommendation G.729 annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *Communications Magazine, IEEE*, vol. 35, pp. 64 –73, Sept 1997.