MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Ultrasonic Doppler Sensor for Voice Activity Detection

Kaustubh Kalgaonkar, Rongquiang Hu, Bhiksha Raj

TR2007-106 August 2008

Abstract

This paper describes a robust voice activity detector using an ultrasonic Doppler sonar device. An ultrasonic beam is incident on the talker's face. Facial movements result in Doppler frequency shifts in the reflected signal, that are sensed by an ultrasonic sensor. Speech-related facial movements result in identifiable patterns in the spectrum of the received signal, that can be used to identify speech activity. These sensors are not affected by even high levels of ambient audio noise. Unlike most other non-acoustic sensors, the device need not be taped to a talker. A simple yet robust method of extracting the voice activity information from the ultrasonic Doppler signal is developed and presented in this paper. The algorithm is seen to be very effective and robust to noise, and can be implemented in real time.

IEEE Signal Processing Letters, Oct. 2007

Ultrasonic Doppler Sensor for Voice Activity Detection

Kaustubh Kalgaonkar†,Rongquiang Hu‡,Bhiksha Raj⋆

***Abstract*— This paper describes a robust voice activity detector using an ultrasonic Doppler sonar device. An ultrasonic beam is incident on the talker's face. Facial movements result in Doppler frequency shifts in the reflected signal, that are sensed by an ultrasonic sensor. Speech-related facial movements result in identifiable patterns in the spectrum of the received signal, that can be used to identify speech activity. These sensors are not affected by even high levels of ambient audio noise. Unlike most other non-acoustic sensors, the device need not be taped to a talker. A simple yet robust method of extracting the voice activity information from the ultrasonic Doppler signal is developed and presented in this paper. The algorithm is seen to be very effective and robust to noise, and can be implemented in real time.**

I. Introduction

VOICE Activity Detectors (VAD) are used to separate regions of speech from non-speech in voice recordings. VADs are important components of speech coding, denoising and recognition systems. VAD algorithms have typically been based on measurements derived from the audio signal itself, such as energy and zero-crossing rates [1], statistical models of speech and noise components of the audio [2], source separation and decision-making based on combination of different features computed from the audio signal [3], etc.The performance of these algorithms often deteriorates rapidly with increasing levels of ambient/background noise.

A recent trend has been the use of *auxiliary* sensors to provide additional evidence of speech activity. Devices like glottal electromagnetic sensors (GEMS) [4], P-mics [5], electroglottographs (EGG) [6] and bone-conduction microphones [7], provide secondary measurements of the speech production process and are relatively insensitive to audio noises. While systems deploying such auxiliary sensors have improved performances over conventional VADs, they have one serious drawback: the auxiliary sensors must be in physical contact with the talker. Bone conduction microphones must be mounted to sense vibrations of facial bones, P-mics, GEMS and EGG sensors must be mounted on the talker's throat.

In a variety of applications such as information kiosks, automotive interfaces, multi-user UIs [8] etc., it isdesirable to have hands-free UIs that can automatically detect when a speaker is addressing them and also endpoint the speech, without requiring direct manipulation by the user. Often these applications are deployed in noisy environments where speech-only VADs may be ineffective. Contact-based secondary sensors like *electroglottograph, bone-conduction mic etc.* are clearly not useful in these applications, as each user who has to use this device (e.g. a kiosk in the mall or bus, train station) will have to mount the sensor on his/her neck or face before using the application.

In [9] Hu and Raj introduced an acoustic-Doppler-based auxiliary sensor for VAD that does not require direct mounting on the talker's person. It consists of an ultrasonic transmitter-receiver pair that is deployed at a distance from the talker and utilizes the Doppler effect to derive information about the movement of the talker's mouth. In addition to being deployable from a distance, the ultrasonic sensor also has the advantage of being very inexpensive (*e.g.* we built an ultrasonic transmitter receiver pair using off-the-shelf components for 3 USD).

Unlike other auxiliary sensors such as GEMS, EGG and P-mics, the measurements derived by the Doppler sensor (which relate the *velocity* of facial components to the *frequency* of the captured signal) are not linearly relatable to speech. Hu et. al. use a support vector machine classifier that combines evidence from the speech signal and the Doppler sensor to classify frames of incoming audio as speech or non-speech. The classifier must be trained offline on joint speech and Doppler recordings, and consequently the performance of the algorithm is highly dependent on the training data used.

In this paper we present a new algorithm for extraction of VAD information from acoustic Doppler readings of the talker's mouth. We utilize a simple FM demodulation scheme to extract combined frequency and energy measurements that are used to determine if the talker is speaking. Unlike the algorithm in [9], no training is required. Further, the algorithm is observed to obtain highly accurate VAD from only the Doppler measurements, without utilizing the acoustic data itself,effectively making the performance of the VAD independent of the background noise level. The algorithm is computationally efficient and can be implemented in real-time.

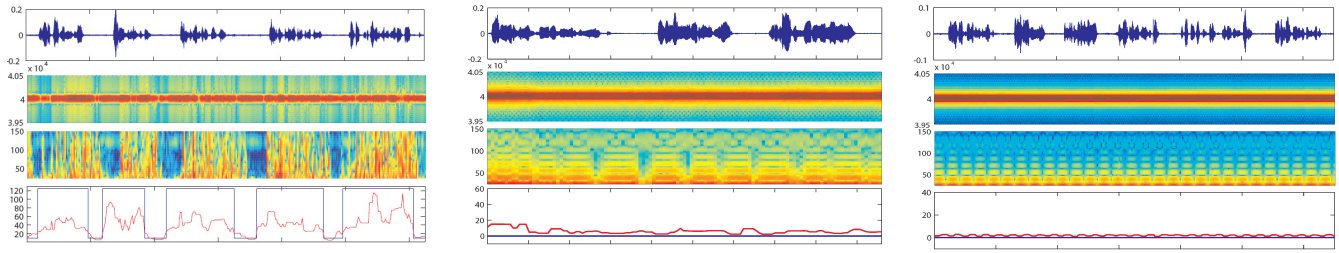
The paper is organized as follows: Section 2 discusses the Doppler effect in the context of voice activity detection and the acoustic Doppler sensor. In Section 3 we describe the demodulation of the Doppler signal. Section 4 explains our VAD algorithm. Section 5 presents experimental evaluation of the proposed VAD algorithm. Finally, conclusions are presented in Section 6.

† School of Electrical and Computer Engineering, Georgia Institute of Technology. Atlanta GA 30332. Email: kaustubh@ece.gatech.edu

‡ Ditech Networks Inc. 825, East Middlefield Rd, Mountain View, CA 94043. Email: rhu@ditechnetworks.com

⋆ Mitsubishi Electric Reserch Lab. Cambridge MA 02139. Telephone: (617) 621 7593. Email: bhiksha@merl.com

This work was performed at Mitsibishi Electric Research Labs



(a) Talker facing the mic/Doppler receiver (b) Paper between talker and mic/Doppler receiver (c) Talker not facing the mic/Doppler receiver

Fig. 1. Audio signal, spectrogram of Doppler signal, spectrogram of demodulated Doppler signal and energy track overlaid with VAD output for three conditions.

II. The Doppler Effect and the Acoustic Doppler Sensor

The Doppler effect is the phenomenon by which the frequency perceived by a listener who is motion relative to a signal emitter is different from that emitted by the source. Specifically, if a signal emitter emits a frequency f that is reflected by an object moving with velocity v with respect to the emitter, the reflected frequency $\hat{f}$ sensed at the emitter is shifted with respect to the original frequency f, and is given by

$$\hat{f} = \frac{v_s + v}{v_s - v} f \approx \left(1 + \frac{2v}{v_s}\right) f \tag{1}$$

where v_s is the velocity of sound in the medium. The approximation to the right in Equation (1) holds true if $v \ll v_s$. If the signal is reflected by multiple objects moving with different velocities, multiple frequencies will be sensed, one from each object.

Fig. 2. Recording setup. The larger sensor in the middle is the microphone

In this paper we utilize an acoustic-Doppler-based sensing device that uses the above principle to sense movements of a talker's face. The device consists of a transmitter that emits a continuous ultrasonic tone at 40 kHz and a transducer that is tuned to receive signals around 40 kHz. The transmitter and receiver are mounted close to the microphone that captures the speech signal. Figure 2 shows an example configuration. In our setup the ultrasonic transmitter and receiver are both about 8 mm in diameter, which is approximately equal to the wavelength of the 40 kHz tone[1]. Both the sensors are relatively directional, with a beamwidth of about 60°.The talker should face the microphone/ultrasonic transmitter arrangement while speaking and must positioned within the beam of the receiver, no more than 60 cm from device. The emitted ultrasonic tone is incident on the talker's face and the reflected signals are sensed by the receiver. The lip movements are not registered if the talker is more than 60 cm from the device, making the setup robust to spurious movements and bystanders beyond the range of 0.6 m.

The human face is an articulated object with multiple components moving with different velocities when a person speaks. The ultrasonic signal reflected by the talker's face therefore has a range of frequencies. These frequencies represent Doppler shifts caused by the velocities of facial components. The orientation of the talker's face affects the spectrum of the reflected ultrasonic signal – the spectra observed when talkers face the device are different from those observed when they do not. Similarly, the spectra of ultrasonic signals reflected from a talker's face are different from those of reflections from other moving or static objects. Figure 1 shows reflections from various targets. The objects and conditions of the experiments were specifically chosen to elaborate the difference in the reflections. Figure 1(a) shows the audio waveform and spectrogram of the corresponding ultrasonic Doppler signal, when the talker faces the sensor. Figures 1(b) an 1(c) show the audio and the Doppler spectrogram when the ultrasonic signal is not reflected directly from the talker's face and when the talker does not face the sensor respectively. The spectrum of the received ultrasonic signal when the talker is facing the device is very different from the other two cases where the ultrasonic reflection are not recording lip movements, demonstrating that there is a strong correlation between the presence of speech and the characteristics of the spectra of the reflected ultrasonic signal.

III. Analysing the Signals from the Doppler Sensor

The Doppler sensor emits a continuous tone that may be represented as $s(t) = \sin(2\pi f_c t)$, where f_c is the emitted frequency (40 kHz in our case). The target (*i.e.* the talker's face) is an articulated object that can be modeled as a discrete combination of moving components, where the i^{th} component has a time-varying velocity $v_i(t)$. The signal sensed at the Doppler(Ultrasonic) receiver is a sum of the signals reflected by various moving components. Note that this is different from the conventional Doppler-sensing scenario where a *single* target moving with a single, possibly time-varying velocity, is to be tracked. Equation (2) represents a situation where several targets are simultaneously sensed. Fortunately in this particular case we do not need to resolve/track individual targets;it is sufficient to detect the presence of these multiple targets. This can be done by processing the combined reflections from the

[1]A more compact setup may be obtained using smaller ultrasonic elements, that are freely available. Also, a single broad-band sensor can replace the combination of the audio microphone and the ultrasonic sensor

face. The combined reflection from all moving components is given by:

$$d(t) = \sum_i a_i \sin\left(2\pi f_c \left(t + \frac{2}{v_s}\int_0^t v_i(\tau)d\tau\right) + \phi_i\right) \quad (2)$$

Equation (2) utilizes the approximate form of the Doppler's equation given in Equation (1). a_i is the amplitude of the signal reflected by the i^{th} component and is related to its distance from the sensor. Although a_i is also time-varying, the changes are relatively slow, compared to the cosine terms. For the purpose of our analysis, we therefore assume it to be a constant gain term. ϕ_i is a phase term representing the relative phase differences between the signals reflected by the various facial components.

If f_c is considered to be the carrier frequency then Equation (2) represents the sum of multiple Frequency Modulated (FM) signals operating on the carrier frequency f_c.

Most of the information related to the movement of facial components resides in the frequency of the signals as seen in Equation (2). For effective VAD, we demodulate the signal using simple frequency demodulaton [10], so that the frequency information is now expressed as the *amplitude* of the cosine components. This demodulated signal also provides a measure of the energy of these movements. Frequency demodulation of the received signal results in a spectral-decomposition like output.

To demodulate the signal, we first differentiate the received signal $d(t)$. Differentiating Equation (2) we get

$$\frac{d}{dt}d(t) = \sum_i 2\pi a_i f_c \left(1 + \frac{2v_i(t)}{v_s}\right) \cdot \cos\left(2\pi f_c \left(1 + \frac{2}{v_s}\int_0^t v_i(\tau)d\tau\right) + \phi_i\right) \quad (3)$$

The derivative of $d(t)$ is amplitude demodulated; by multiplication with a sinusoid of frequency f_c, followed by low-pass filtering with cut-off frequency of f_c. This gives us

$$\text{LPF}\left(\sin(2\pi f_c t)\frac{d}{dt}d(t)\right) = -\sum_i 2\pi a_i f_c \left(1 + \frac{2v_i(t)}{v_s}\right)\sin\left(\frac{2\pi f_c}{v_s}\int_0^t v_i(\tau)d\tau + \phi_i\right) \quad (4)$$

where LPF represents the low-pass-filtering operation.The signal represented by Equation (4) encodes velocity terms in both, the amplitudes and frequencies of its spectral components. If the signal is analyzed in short analysis windows, the velocities of the frequencies do not change significantly within the analysis frame and the right hand side of Equation (4) can be interpreted as a frequency decomposition of the left hand side. The signal contains energy primarily at frequencies related to the various velocities of the moving facial structures. The energy at any velocity is a function of the number and distance of facial components moving with that velocity, as well as the velocity itself.

Figure 1(a) shows the spectrogram of the original Doppler signal as well as that of the demodulated signal (given by Equation (4)). The latter exhibits greater visual correlation to the presence of speech than the former.

In general, speech-related facial movements in the direction of the Doppler sensor are relatively slow. Correspondingly, most of the speech-related energy in the spectrum of the demodulated Doppler signal is found to lie in the 50 Hz-125 Hz frequency range (this was also verified by processing hours of speech and doppler data). Frequencies outside this range, although related to speech activity, are often corrupted by the carrier frequency, as well as harmonics of the speech signal including any background speech or babble, particularly in voiced segments. Hence, we restrict our analysis to the 50 Hz-125 Hz frequency band for VAD. Figure 1 shows the 25 Hz-150 Hz band of the spectrograms of the demodulated Doppler signal for three recording conditions. The spectrogram of the reflected ultrasonic signal where the talker is facing the sensor distinctly stands out.

IV. Voice Activity Detection

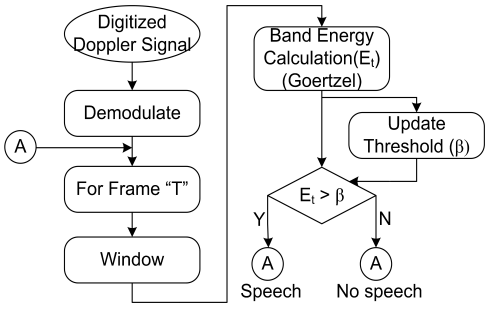


Fig. 3. Flow Chart of VAD Algorithm

The algorithm for detecting speech activity is shown by the flowchart in Figure 3. The Doppler signal is first digitized and demodulated. Since the data are digitized, the derivative of Equation (4) is obtained through Euler approximation. The demodulated signal is segmented into frames of 64 ms. Adjacent frames overlap by 50%.

A hamming window is applied to each of the frames, following which the energy in the 50 Hz-120 Hz frequency band is extracted using Görtzel's algorithm [11].

The energy contour is median filtered to reduce the high fluctuations within the speech regions. Finally, to determine if the t^{th} frame contains speech activity, the median filtered energy E_t in the frame is compared to an adaptive threshold β_t. The threshold β_t is adapted to track the background level of the Doppler signal as follows $\beta_t = \beta_{t-1} + \mu(E_t - E_{t-1})$, where μ is an adaptation factor that can be adjusted for optimal performance. Figure 1 also shows the plot of the energy extracted from the Doppler signal overlaid by a plot of the regions determined to be speech by our VAD algorithm. The energy contour has prominent peaks that correspond to regions of speech only when the talker is facing the sensor. Consequently, the VAD algorithm only detects presence of speech in 1(a). Even though speech is present and recorder on the audio channel, for cases 2 and 3 Figures 1(b) and 1(c), the VAD output is negative due to the absence of the energy in the band of interest of the received Doppler signal.

V. Results

A small corpus of simultaneous recording of speech and Doppler sonar signals was made at Mitsubishi Electric Re-

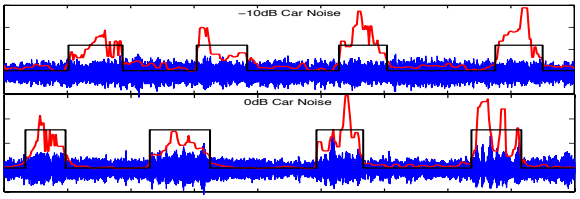


Fig. 4. Audio, energy contour and Doppler VAD output at 0 dB and -10 dB

search Labs. The corpus consists of 4 talkers(3 male, 1 female) speaking 30 TIMIT sentences under different conditions: quiet, car, babble, competing speech and music. The noise was not digitally added, i.e. the recordings were made in the presence of these noise sources. The boundaries of the speech signal were hand labeled (to provide the "ground truth" for the VAD). The SNR of the utterances was estimated from the RMS values of the speech and non speech regions. SNR was varied over a large region (-10 dB to 20 dB). Two voice

TABLE I
ACCURACY OF VOICE ACTIVITY DETECTORS

Noise Type	SNR	Audio only VAD (%)	Doppler VAD(%)
Office	-10dB	8.01	92.8
	0	89.55	95.11
	10	90.47	97.43
Car	-10dB	1	92.92
	0dB	54.84	97.68
	10dB	67.45	96.23
Babble	-10dB	5	94.42
	0dB	52.50	96.05
	10dB	60.76	95.17
Speech	-10dB	0	93.94
	0dB	57.02	97.37
	10dB	62.78	96.84
Music	-10dB	2	95.56
	0dB	50.89	95.52
	10dB	54.32	96.21

activity detectors were implemented, one was based on the audio recordings (speech) only and the other was the Doppler-based algorithm described in Section 4. The audio-only voice detector uses a prior speech presence probability model with minimum statistics noise estimation [2]. Table I shows the frame-wise percentage VAD accuracy. Doppler based VAD is very robust to noise and has nearly constant performance at all the noise levels. The performance of the audio-only VAD deteriorates rapidly with increase in noise power.

Also, the Doppler based VAD is fairly immune to noise type and the detection accuracy in the presence of competing speech is comparable to that in the presence of any other noise. On the other hand, the performance of the audio-only VAD deteriorates in the presence of competing speech as the noise and actual speech signal have the same characteristics. Figure 4 shows the VAD applied to audio signals recorded at 0 dB and -10 dB SNR. Audio-energy based VAD will not be able to isolate speech in such conditions, but the Doppler-based voice activity detector was able to identify each speech segment correctly.

Since the Doppler-based VAD detects speech through facial motion, it is likely that it will be triggered by other kinds of motion as well. As a test, we let both the speech-based and Doppler-based detectors run continuously in an open office space, with people moving around, background speech, and ambient noise. The "Doppler-mic"shown in Figure 2 is mounted on microphone stand 5 ft above the ground ensuring that a 40 KHz tone could be directed towards a talkers face. Recordings were made for over 80 minutes. Nobody addressed the system in this time. The speech energy based speech detector generated 30 spurious frame-level voice activity decisions per minute on average in this period, as compared to 1 false decision per minute generated by the Doppler-based VAD.

VI. CONCLUSION

The proposed voice activity detector is observed to be very robust in all type and levels of noise. The Doppler sensor provides complimentary data that is not captured by the acoustic microphone: it captures facial movement. The receiver is relatively insensitive to audio noise, by virtue of which the Doppler-based VAD can isolate regions of speech activity in extremely high noise conditions.

The algorithm currently does not utilize the audio signal itself to determine speech activity, since our goal was to enable robust speech activity detection under noise conditions where the audio signal becomes an unreliable indicator of speech activity. One consequence of this is that the VAD can sometimes be triggered by spurious lip movements or other similar motion. Such false alarms can be greatly reduced by appropriately correlating results obtained from the Doppler VAD to the audio channel, in a manner that takes the current noise level into account. Further, by incorporating the energy features used in this paper in a learning framework such as [9], the VAD performance may be further improved.

While the current paper deals only with VAD, the correlation between the Doppler measurements and the underlying speech may be utilized for improved denoising of the speech signal. These and related topics are topics of current and future investigation.

REFERENCES

[1] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *The Bell System Technical Journal*, vol. 54(2), pp. 297–315, 1975.
[2] Saeed Gazor and Wei Zhang, "A soft voice activity detector based on a laplacian-gaussian model," *IEEE Trans. on Speech and Audio Process*, vol. 11, pp. 498505, 2003.
[3] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," *IEEE Trans. Acoust. Signal Speech Process*, vol. 8, 2000.
[4] G.C. Burnett and et. al., "The use of glottal electromagnetic micropower sensors (gems) in determining a voiced excitation function," *138^{th} Meeting of the Acoustical Society of America*, 1999.
[5] M.V. Scanlon, "Acoustic sensor for health status monitoring," *Proceedings of IRIS Acoustic and Seismic Sensing*, vol. 2, pp. 205–222, 1998.
[6] M. Rothenberg, "A multichannel electroglottograph," *Journal of Voice*, vol. 6, pp. 36–43, 1992.
[7] J. Hershey and et. al., "Model-based fusion of bone and air sensors for speech enhancement and robust speech recognition," *ISCA SAPA*, 2004.
[8] P. H. Dietz and D. L. Leigh, "Diamondtouch: A multi-user touch technology," *ACM Symposium on UIST*, pp. 219–226, 2001.
[9] Rongqiang Hu and Bhiksha Raj, "A robust voice activity detector using an acoustical doppler radar," *IEEE ASRU 2005*, pp. 319 –324, 2000.
[10] Simon Haykin, *Communication Systems*, Wiley, 2000.
[11] A. V. Oppenheim and R. W. Schafer, *Discrete time signal processing*, Prentice-Hall, 1990.