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

A Robust Voice Activity Detector Using an Acoustic Doppler Radar

Rongqiang Hu, Bhiksha Raj

TR2005-159 November 2005

Abstract

This paper describes a robust voice activity detector using an acoustic Doppler radar device. The sensor is used to detect the dynamic status of the speaker's mouth. At the frequencies of operation, background noises are largely attenuated, rendering the device robust to external acoustic noises in most operating conditions. Unlike the other non-acoustic sensors, the device need not be taped to the speaker, making it more acceptable in most situations. In this paper, various fetures computed from the sensor output are exploited for voice activity detection. The best set of features is selected based on robustness analysis. A support vector machine classifier is used to make the final speech/non-speech decision. Experimental results show that the proposed doppler-based voice activity detector improves speech/non-speech classification accuracy over that obtained using speech alone. The most significant improvements happen in low signal-to-noise (SNR) environments.

IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)

A ROBUST VOICE ACTIVITY DETECTOR USING AN ACOUSTIC DOPPLER RADAR

Rongqiang Hu [1], *Bhiksha Raj* [2]

[1]Georgia Institute of Technology, [2]Mitsubishi Electric Research Laboratories

ABSTRACT

This paper describes a robust voice activity detector using an acoustic Doppler radar device. The sensor is used to detect the dynamic status of the speaker's mouth. At the frequencies of operation, background noises are largely attenuated, rendering the device robust to external acoustic noises in most operating conditions. Unlike the other non-acoustic sensors, the device need not be taped to the speaker, making it more acceptable in most situations. In this paper, various features computed from the sensor output are exploited for voice activity detection. The best set of features is selected based on robustness analysis. A support vector machine classifier is used to make the final speech/non-speech decision. Experimental results show that the proposed doppler-based voice activity detector improves speech/non-speech classification accuracy over that obtained using speech alone. The most significant improvements happen in low signal-to-noise (SNR) environments.

1. INTRODUCTION

Voice activity detectors (VAD) are used to demarcate regions of conversational speech from silent or non-speech regions of a speech signal. VADs are important to many speech processing applications such as speech enhancement, speech coding, speech recognition etc. Various VAD algorithms have been proposed in the literature, that are based on zero crossing rates, spectral representatives (LPC, LSF, etc.), statistical speech and noise modeling [1], source separation, and decision-making based on a combination of different features [2]. The algorithms perform well in quiet or high SNR environments. But the performance drops dramatically as the level of background noise increases.

Conventional voice activity detectors work chiefly from measurements obtained from the speech signal. A recent trend has been the use of measurements from *secondary* sensors in additional to the primary speech recording, for the measurement of speech signals in the presence of strong background noise. These sensors typically provide measurements of one or more aspects of the speech production process such as a coarse measurement of the speech signal itself, or measurements of glottal activity, as a proxy for the actual speech and tend to be relatively immune to acoustic noise. These sensors typically do not provide enough information about the speech generation process to replace microphone sensors; instead, these sensors must be used in conjunction with a microphone and additional signal processing in order to augment the acoustic speech signal for the purpose of speech enhancement, coding and recognition in high-noise environments. Secondary sensors have been shown to greatly improve the performance of voice activity detection in high noise environments.

Most current secondary sensors used for voice activity detection, however, suffer the drawback that they require contact with the speaker. Bone conduction microphones must be mounted on a the jaw bone. Physiological microphones (P-mics), throat microphones and the non-acoustic glottal electromagnetic sensors (GEMS) must all be mounted on the speaker's face or throat. This restricts their utility in most applications.

In this paper we propose the use of an entirely different variety of secondary sensor for voice activity detection - a Doppler acoustic radar. The Doppler radar consists of a high-frequency ultra sound emitter and an acoustic transducer that is tuned to the transmitted frequency. The ultra-sound tone emitted from the sensor is reflected from the speaker's face and undergoes a Doppler frequency shift that is proportional to normal velocity of the portion of the face that it is reflected from. The spectrum of the reflected signal thus contains an spectrum of frequencies that represent the motion of the speakers cheeks, lips, tongue, etc. The voicing state of the speaker (i.e. speech. vs. non-speech activity) is estimated using a support vector machine classifier on appropriate measurement derived from this reflected signal.

While the Doppler measurements are not as detailed as those obtained from secondary sensors such as P-mics or GEMS sensors, the measurements obtained from it are nevertheless adequate for voice activity detection. Experiments conducted on spoken utterances collected in the presence of a variety of background noises show that the proposed VAD algorithm based on acoustic Doppler measurements results in significantly better voice activity detection than that obtained from measurements of the speech signal alone. Additionally the proposed secondary sensor has the advantage that it need not be mounted on the speaker. In fact it is effective even at a distance of 10-15cm from the speaker. It is also far more economical than cameras (which can also be used to derive useful secondary measurements from a distance) - an acoustic Doppler radar setup can be constructed for less than $10.

The rest of the paper is arranged as follows: in Section 2 we briefly review the problem of voice activity detection, and the use of secondary sensors for the purpose. In Section 3 we describe the acoustic Doppler radar based secondary sensor. In Section 4 we present an analysis of the mutual information between the signal captured by the proposed Doppler sensor and the speech signal. In Section 5 we describe the features computed from the Doppler signal. In Section 6 we review the Support Vector Machine classifier used for speech/non-speech detection. In Section 7 we describe our experimental evaluation of the proposed voice activity detection algorithm and finally in Section 8 we present our conclusions.

2. VOICE ACTIVITY DETECTION USING SECONDARY SENSORS

Voice activity detection is the problem of determining whether any segment of a speech recording occurs within a continuously spoken utterance or if it actually represents the bracketing non-speech regions. This has traditionally been performed using the recorded

speech signal itself. When the speaker is speaking, the recorded signal $Y(f)$ (as represented in the frequency domain) is a mixture of speech $S(f)$ and noise $N(f)$, i.e. $Y(f) \cong S(f) + N(f)$. When no speech is uttered, the sensor captures chiefly noise, i.e. $Y(f) = N(F)$. The goal of VAD is to determine whether speech is present or not from observations of $Y(f)$.

The simplest VAD procedures are based on thresholding of measurements such as zero crossings and energy. More sophisticated techniques (e.g. [1]) employ statistical models applied either to the signal itself, or to features derived from it, such as spectra, LPC residuals, etc. These algorithms perform very well in clean and low-noise environments. However, in real-world environments with high levels of noise they often perform poorly.

The use of secondary sensors to improve the noise robustness of VAD has become increasingly popular in recent times. These are sensors that obtain secondary measurements either of the speech signal, or of the underlying speech generation process. An important criterion for an effective secondary sensor is that its measurements must be relatively immune to or independent of the background noise that affects the speech signal itself.

Most current research on secondary sensors for VAD is concentrated on sensors whose measurements are linearly relatable to the speech signal. From a speech production perspective, the speech signal can be modeled as

$$S(f) = G(f)V(f)R(f) \tag{1}$$

where $G(f)$, $V(f)$, and $R(f)$ represent glottal excitation, the frequency response of the vocal cavity and lip radiation respectively. In most current research, measurements from the secondary sensor are required to be linearly relatable to one or more of the components on the right hand side of Equation 1. That is, the measurements must be of the form $Y(f) = H(f)S(f)$ in speech regions, where $H(f)$ represents a linear filter. Additionally, and more importantly, they must be relatively insensitive to the noise that corrupts the speech signal, i.e. in non-speech regions $Y(f) \ll N(f)$.

A variety of secondary sensors have been proposed that satisfy these conditions. Examples of such sensors are the physiological microphone (P-mic), which measures the movement of the tissue near the speaker's throat, and the bone-conduction microphone, which measures the vibration of bone associated with speech production. In these sensors $H(f)$ is a low-pass filter. The signal captured in non-speech areas is significantly lower than $N(f)$. A second kind of secondary sensor seeks to provide a function of the glottal excitation, e.g. the Electroglottograph (EGG)[3], and the glottal electromagnetic sensor (GEMS) [4]. In this case, $X(f) \cong G(f)$ during voiced speech, and the corrupting noise is nearly 0 in non-speech regions.

All of these secondary sensors have shown promise in many speech applications, such as voice activity detection, speech enhancement and coding [5, 6, 7]. However, they typically require that sensors be placed in direct contact with the talkers skin in a suitable location, making them uncomfortable to users. Also, the measurements they provide are not always perfectly linearly relatable to speech. While the P-mic and bone-conduction microphone provide relatively noise-free measurements at low-frequencies, they do not capture speech-related information in higher frequencies, and are unreliable in unvoiced regions. The EGG and GEMS sensors approximate glottal excitation function during voiced speech, but they can not provide any measurement about unvoiced speech. The high cost of these sensors also makes them impractical in many applications.

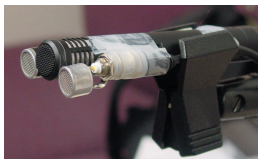

Fig. 1. The Doppler-augmented microphone used in our experiments. The two devices taped to the sides of the central audio microphone are a high-frequency emitter and a high-frequency sensor.

3. THE ACOUSTIC DOPPLER SENSOR

Contrary to current secondary sensors, the acoustic Doppler radar that we propose to use as a secondary sensor does not attempt to obtain measurements that are linearly relatable to the speech. Instead, it is based on a very simple principle: the facial structures of a person's face, including their cheeks, lips, and particularly their tongue move when the person speaks[1]. It should hence be possible to determine if the person is speaking or not simply by observing whether they are moving their vocal apparatus or not. While such a determination can be made using visual aids such as a camera, these solutions tend to be expensive, both in economical and computational terms. A simpler solution might be use a simple motion detector; however simple detectors cannot distinguish between the range of motions that a speaker's vocal apparatus can make. Such measurements can, however be made by an Doppler radar.

Acoustic Doppler radars are based on a simple principle: when a high-frequency tone is reflected from a moving object, the reflected signal from the object undergoes a frequency shift that is related to the velocity of the object in the direction of the radar. If the tone emitted by the radar has a frequency f and the velocity of the object in the direction of the radar is v, the frequency of the reflected signal f' is related to f and v by

$$f' = \frac{(c+v)f}{(c-v)} \tag{2}$$

where c is the velocity of sound. When the target object has several moving parts, such as a mouth, where each part has a different velocity, the signal reflected by each component of the object has a different frequency. The reflected signal captured by the radar therefore has an entire spectrum of frequencies that represent the spectrum of velocities in the moving parts of the target.

When the target of the acoustic Doppler radar is the human mouth and its surrounding tissue, the spectrum of the reflected signal represents the set of velocities of all moving parts in the mouth, including the cheeks, lips and tongue. In addition, the energy in the reflected signal depends on the configuration of mouth, e.g. the signal reflected from an open mouth has less energy due to the absorbtion of the back of the mouth or, if the radar is placed at an angle, due to the fact that some of the incident signal travels straight through unimpeded (and is reflected perhaps by a relatively distant object with significant attenuation).

Figure 1 shows the acoustic Doppler radar augmented microphone that we have used in our work. In this embodiment, the complete setup has three components. The central component is a conventional acoustic microphone. To one side of it is a ultra-

[1] We do not account for special cases such as closed-mouth talking and ventriloquist speech in this assumption

sound emitter that emits a 40Khz tone. To the other side is a high-frequency transducer that is tuned to capture signals around 40Khz. The microphone and transmitter are well-aligned, and placed directly pointed to the mouth. The dynamic status of the mouth moving is measured by the device. It must be noted that the device also captures high-frequency harmonics from the speech and any background noise; however these are significantly attenuated with respect to the level of the reflected Doppler signal in most standard operating conditions[2]. The device does not require contact with the skin. As may be inferred from Figure 1, the acoustic Doppler was placed at exactly the same distance as the desktop microphone itself from the speakers, in our experiments. The cost of the entire setup shown in the Figure is not significantly greater than that of the acoustic microphone itself: the high-frequency transmitter and receiver both cost less than a dollar. The transmission and capture of the Doppler signal can be performed concurrently with that of the acoustic signal by a standard stereo sound card. Since the high-frequency transducer is highly tuned and has a bandwidth of only about 4Khz, the principle of band-pass sampling may be applied, and the signal need not be sampled at more than 12Khz (although in our experiments we have sampled the signal at 96Khz).

4. MUTUAL INFORMATION ANALYSIS OF THE DOPPLER SENSOR

In order to be effective, the measurements from the acoustic Doppler sensor must be related to the underlying clean speech signal. Stated otherwise, knowledge of the Doppler signal must reduce the uncertainty in our knowledge of the speech signal. The predictability of the speech signal from the Doppler measurement can be stated as the mutual information between the two signals.

The mutual information (MI, $I(x,y)$) between two variables x and y is described as

$$I(x,y) = D[P(x,y) \| P(x), P(y)] \quad (3)$$

$$= \int_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} dxdy \quad (4)$$

where $P(x)$ and $P(y)$ are the densities of x and y respectively, and $P(x,y)$ is the joint density of x and y. D denotes the Kullback-Leibler divergence, also known as the relative entropy. The MI covers all kinds of linear and non-linear dependencies [8]. In case the statistical distributions of the variables are unknown and only a limited amount of samples of the variables are available for measurement, a non-parametric estimator is proposed in [9]. The algorithm approximates the mutual information arbitrarily closely in probability by calculating relative frequencies on appropriate partitions of the data space and achieving conditional independence on the rectangles of which the partitions are made.

Reviewing the objectives of employing a secondary sensor in robust speech processing , the qualification of a secondary sensor in robust speech processing can be summarized in information theoretic terms as follows:

- High dependency between the outputs of the secondary sensor X and clean speech S, i.e. $I(X,S)$ is large.
- High independence of the outputs of a secondary sensor X and noise N,i.e. $I(X,N)$ is low.

In recordings obtained from high-noise environments, the second condition may also be stated as a requirement of low $I(X,Y)$, i.e. of independence between the doppler and noisy speech measurements. Given these criteria, the robustness of a secondary sensor can be represented as the normalized change of mutual information in noisy environments.

$$\Delta I(X,Y\|SNR) = \frac{I(X,S) - I(X,Y\|SNR)}{I(X,S)} \quad (5)$$

The greater the value of $\Delta I(X,Y\|SNR)$ the more useful the measurements of the sensor can be expected to be in processing highly noisy speech.

The MI analysis of recordings from GEMS, P-mic and EGG sensors is listed in Table 1. The results confirm the observations in [6, 7] that GEMS contains more secondary information about speech than P-mics and EGG, and is also more robust than the others two. As described in section 2, P-mic recordings contain some level of acoustic noise. All of these sensors have been applied to robust speech processing and have produced improved performance in voice activity detection and speech enhancement [7].

Table 1. Mutual Information between the sensor outputs and acoustic signals

Clean Environment	GEMS	P-mic	EGG
I(X,S)	0.272	0.075	0.091

Noise	Office	Tank	Shoot	Helicopter
$\Delta I(X,Y)$	(23dB)	(1dB)	(13dB)	(3dB)
GEMS	0.202	0.993	0.743	0.996
P-Mic	0.280	0.693	0.027	0.640
EGG	0.044	0.912	0.626	0.967

The MI between the Doppler radar and acoustic speech signals is given in Table 2. The table analyses signals captured in the presence of both stationary and non-stationary noises. The similarity between the numbers in Tables 1 and 2 indicate that the Doppler radar sensor can provide effective secondary information for robust speech processing in noisy environments.

Table 2. Mutual Information between the Doppler radar outputs and acoustic signals

Clean Environment	Doppler Radar
I(X,S)	0.097

Noise	Office	Car	Babble	Speech	Music
	(22dB)	(4dB)	(5dB)	(7dB)	(5dB)
$\Delta I(X,Y)$	0.041	0.938	0.959	0.680	0.835

5. FEATURE SELECTION

The motion of the mouth plays an essential role in speech production. In order to produce different sounds, a person must change

[2]The system will however not work if there are any devices in the vicinity that specifically emit noise at 40Khz.

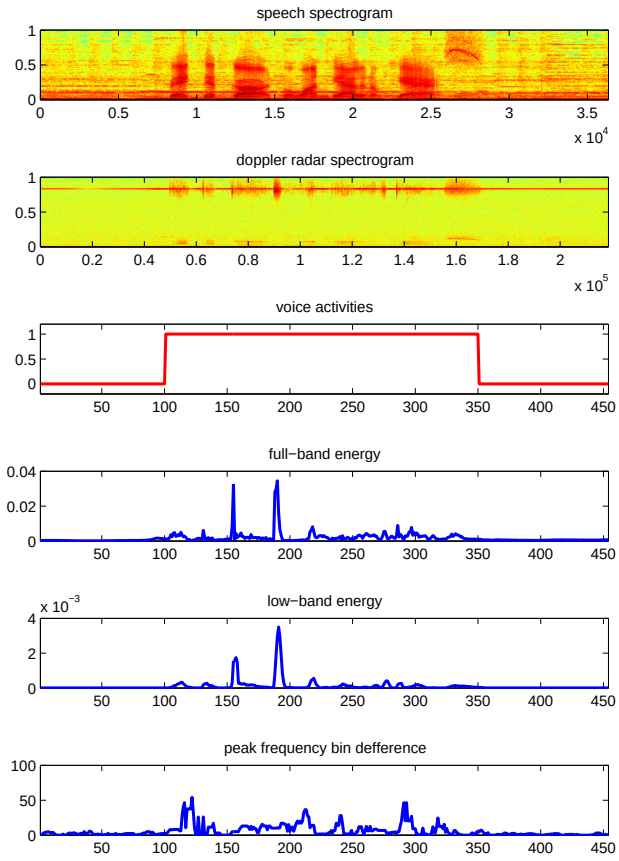

Fig. 2. Example of features in music noise (SNR=12dB)

the configuration of their entire vocal apparatus, particularly the mouth. This is true regardless of whether the sound is voiced or unvoiced. Sensors that measure voicing-based information from the vocal tract are only effective in detecting voiced speech. For unvoiced sounds, such as unvoiced consonants, these sensors do not provide any information. Measuring the dynamic status of the mouth, however, is effective in detecting both voiced and unvoiced sounds.

5.1. Parameter Extraction

In processing Doppler radar signals, two key features are considered: reflected energy and peak reflected frequency. When a tone of frequency f is reflected off a slowly moving object with velocity $v \ll c$, the reflected frequency f' is given by the following modification of Equation 2:

$$f' = \left(1 + \frac{2v}{c}\right) f \tag{6}$$

When a speaker's mouth is closed, there is no motion of the mouth, i.e. $v = 0$. Hence, the observed frequency $f' = f$. While speaking, the mouth and tongue of the speaker move. The velocity of the moving parts of the mouth are typically in the order of 0.1m/s, although the peak velocity of the tongue can be significantly greater. Since our acoustic Doppler device emits a signal at 40KHz, the reflected frequency f' will be in the neighbourhood

Table 3. Mutual Information of features with voice activity labels

Clean	E_f	ΔE_f	E_l	ΔE_l	F_p	ΔF_p
$I(A,L)$	0.878	1.594	1.175	1.932	0.313	2.709

Feature	Office	Car	Babble	Speech	Music
$\Delta I(A,L)$	(22dB)	(4dB)	(5dB)	(7dB)	(5dB)
E_f	0.001	0.012	0.009	0.042	0.006
ΔE_f	0.003	0.010	0.022	0.013	0.029
E_l	0.002	0.007	0.016	0.038	0.056
ΔE_l	0.005	0.019	0.039	0.051	0.078
F_p	0.028	0.037	0.055	0.083	0.062
ΔF_p	0.009	0.034	0.026	0.009	0.017

of 40020Hz. Although an entire spectrum of frequencies is reflected from the various moving parts of the mouth, typically one frequency dominates the rest. The actual observed peak frequency can be calculated picking the highest peak from the Fourier transform of radar signals. The velocity of the vocal parts, and therefore the observed peak frequency, vary significantly in time. Therefore, a very high resolution FFT is required in the frequency region (39900Hz-40100Hz) to calculate the accurate peak frequency, in order to distinguish between different states of oral motion.

When the mouth is open, radar signals reach the "walls" of the mouth at various angles. The signals are reflected in many directions. Therefore, the received radar energy varies. This feature can also be used to indicate the speaking status. Since there is, in actuality, an entire range of velocities in the vocal apparatus, the "interesting" signals exist in a frequency range of 39900-40100Hz. We therefore calculate the signal energy in this frequency band as a feature, which we denote as "full-band energy".

In addition to the Doppler reflections, the high-frequency transducer also captures high-frequency harmonics of the speech signal, albeit at highly attenuated levels. Since this information is present, we also choose to use it for voice activity detection. We therefore compute the signal energy in the frequency band (20000Hz-39900Hz) and designate it as "low-band energy".

In addition to these basic features, we also compute difference features that measure their deviation with time. Thus the following set of parameters is extracted from the input doppler signal. These measurements are obtained once every 10ms, and are derived over a 100ms analysis window.

- Peak Frequency (F_p)
- The Peak Frequency Difference (ΔF_p)
- Full-Band Energy (E_f)
- Low-Band Energy (E_l)
- The Full-Band Energy Difference (ΔE_f)
- The Low-Band Energy Difference (ΔE_l)

These features are independent of acoustic disturbance, thus immune to background noise. Figure 2 shows examples of the selected feature, obtained from a signal recorded in noisy conditions.

5.2. Robustness Analysis

The robustness and estimation accuracy are the two most important considerations for selecting features to detect voice activity. As mentioned in the previous sections, mutual information is a useful tool to determine the dependency of two signals.

The dependency of the feature (A) and the corresponding voice activity labels (L) is investigated using the mutual information ($I(A, L)$). The MI gives an indication of estimation accuracy. The variation of MI in different environments ($\Delta I(A, L)$) measures the robustness of a feature. A lower value indicates greater robustness. Table 3 lists the MI results between the extracted features and voice activity labels in a variety of conditions.

From the results, we can conclude that the selected features are very robust to background noise. Each of them will contributes to the voice activity detection. Of the set, the peak frequency bin difference is the most effective feature.

6. SVM CLASSIFIER

We perform the actual speech/non-speech classification of each analysis frame of the signal using a support vector machine (SVM) classifier. Support vector machines are known to provide good classification performance in real-world classification problems which typically involve data that can only be separated using a nonlinear decision surfaces [10, 11].

We use the kernel based variant of the SVM classifier, for which the decision function has the form

$$f(x) = \sum_{i=1}^{N} \alpha_i d_i K(x, x_i) + b \tag{7}$$

where N is the number of support vectors, and $K(x, x_i)$ is the kernel function. $K(x, x_i)$, in this implementation, is a radial basis function (RBF).

$$K(x, x_i) = \exp\{-\Psi(|x - x_i|^2)\} \tag{8}$$

Since the voice activity detection is a binary decision classification problem, a soft margin classifier can be used to address the problem of nonseparable data. Slack variables [10] are used to relax the separation constraints:

$$\begin{cases} x_i \bullet w + b \geq +1 - \xi_i, & for\ d_i = +1 \\ x_i \bullet w + b \leq -1 - \xi_i, & for\ d_i = -1 \\ \xi_i \geq 0, \forall i \end{cases} \tag{9}$$

where d_i are the class assignments, w represents the weight vector defining the classifier, b is a bias term, and ξ_i are the slack variables.

In our implementation, the support vectors are trained using features extracted from a training set with hand-labeled voice activity index. The binary class associated with each analysis frame is the corresponding voice activity index.

7. EVALUATION

A small corpus of simultaneous speech and acoustic Doppler radar signals was recorded at Mitsubishi Electric Research Labs. The corpus includes two speakers speaking 30 TIMIT sentences under five different noise environments: office, car, babble, competing speech, and music. All signals were recorded in the presence of background noise, i.e. the noise was not digitally added. The boundaries of speech were hand labelled and the SNR was estimated from the RMS signal values in the speech and non-speech regions. The SNRs vary in a large range (-5dB to 30dB)

Two voice activity detectors were implemented based on the acoustic speech signals only. One was the prior speech presence probability model with minimum statistics noise estimation, which has been shown to be effective in preserving weak speech cues. The other is an SVM classifier trained on the same database. We simplified this model using speech energy features only derived from a bank of four Mel-scaled filters. A preliminary SVM classifier combining the features from Doppler radar and speech was also tested.

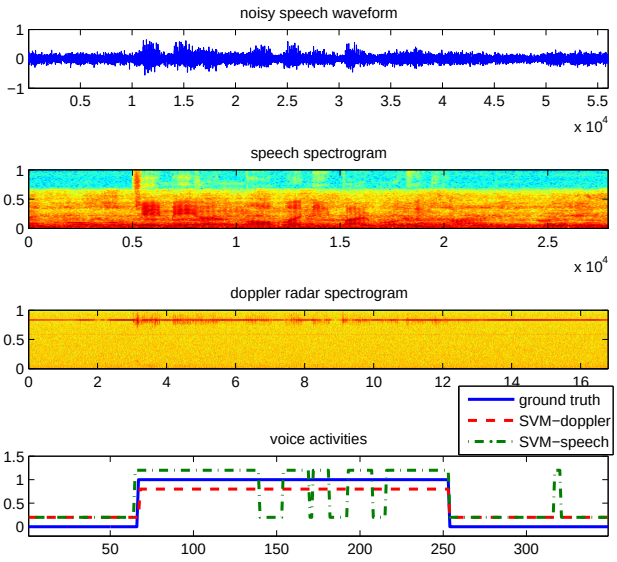

Fig. 3. Illustration of voice activity detection in babble noise (SNR=0dB)

Figure 3 and Figure 4 demonstrate the behavior of two SVM voice activity detectors based on the feature computed from the output of the Doppler radar and speech respectively. The accuracy results in a variety of noise environments are provided in Table 4.

Table 4 shows the frame-wise percentage accuracy of speech/non-speech classification on speech corrupted to varying levels by various noises. We observe that the SVM classifier, based on the features of the Doppler signal, is very robust in noisy environments, outperforming VAD classification based on speech alone in most cases. The robustness of the Doppler based VAD is apparent in that its performance degrades much more slowly with increasing noise than VAD that is based on speech alone.

However in other situations the Doppler-based VAD is not as accurate as that based on speech. The reason for this is simple - people often move their mouths before they begin speaking, and will move their mouths and faces under other conditions as well. Also, the face and vocal apparatus remains relatively stationary during long vowels, giving the impression of vocal inactivity. In such situations, the Doppler radar by itself cannot determine if speech is present. However, in many of these situations cues to the presence of speech are available from the audio signal. Thus, it may be expected that VAD performance can be further improved if the Doppler measurements could be combined with those from the speech signal, for voice activity detection. This hypothesis is borne out by the results obtained when features from the speech and Doppler signals were combined in the SVM classifier: Table 4 shows that VAD performance obtained with a combination of Doppler and speech signals is consistently superior to that obtained with either of the two signals alone.

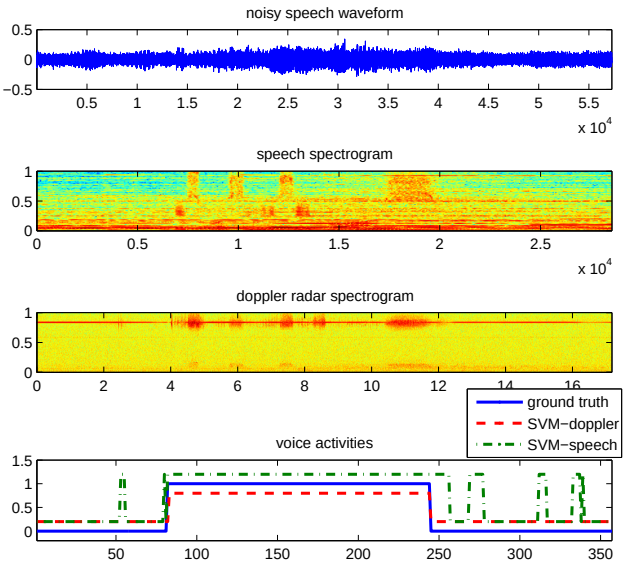

Fig. 4. Illustration of voice activity detection in music noise (SNR=0dB)

Table 4. Accuracy of voice activity detectors

Noise		Min. Stat	SVM	SVM	SVM
Type	SNR	model	Speech	Radar	Comb.
Office	0dB	89.55	95.10	90.78	95.48
	10dB	90.47	95.47	90.37	95.23
	20dB	93.81	96.49	92.63	97.72
	AVG.	92.24	95.97	91.68	96.01
Car	0dB	54.84	88.32	92.69	91.72
	10dB	67.45	90.21	92.22	93.01
	20dB	83.58	93.45	90.77	95.03
	AVG.	70.32	91.96	92.01	93.26
Babble	0dB	51.50	65.54	89.28	89.93
	10dB	60.76	73.89	90.76	90.80
	20dB	73.04	88.96	93.29	94.01
	AVG.	65.84	78.90	91.95	92.43
Speech	0dB	57.02	57.53	93.37	91.59
	10dB	62.72	73.34	92.99	92.86
	20dB	74.67	85.97	93.74	94.86
	AVG.	67.11	77.69	93.36	93.27
Music	0dB	50.89	70.39	90.27	89.51
	10dB	54.32	77.92	93.13	93.89
	20dB	63.39	86.23	92.74	94.12
	AVG.	59.20	78.77	92.63	92.73

8. CONCLUSION

The proposed Doppler-radar-based VAD algorithm was observed to be very robust in all noisy environments, particularly when Doppler measurements were combined with measurements of the speech signal. Dramatic improvements are seen particularly in low SNR conditions. The proposed Doppler radar based sensor thus promises to be a highly effective secondary sensor for voice activity detection.

The proposed acoustic Doppler radar provides data about the motion of the face - a measurement that is not directly obtainable from the speech signal itself. The information it provides is thus complementary to that obtainable from the speech signal. Hence, it may be expected that even if the basic speech signal based VAD algorithm were to be improved significantly, its performance could be further enhanced by combining it with the Doppler measurements. Additionally, the Doppler measurements may be complementary to current secondary sensors such as GEMS and bone conduction sensors, and their performance may also be further improved by combining them with the Doppler sensor.

Finally, we have thus far only attempted to use the Doppler measurements to improve voice activity detection. It stands to reason that the improved voice activity detection can be translated to improved signal enhancement as well. Doppler radar measurements may also be useful secondary features for automatic speech recognition. We will address these issues in future research.

9. REFERENCES

[1] Saeed Gazor and Wei Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 498–505, Sept. 2003.

[2] S. G. Tanyer and H. Ozer, "Voice activity detection in non-stationary noise," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 478–482, July 2000.

[3] R. Baken, "Electroglottography," in *J. Voice*, 1992, pp. 98–110.

[4] G. C. Burnett, J. F. Holzrichter, T. J. Gable, and L. C. Ng, "The use of glottal electromagnetic micropower sensors (GEMS) in determining a voiced excitation function," presented at the 138th Meeting of the Acoustical Society of America, Columbus, Ohio, Nov. 1999.

[5] L.C. Ng, G.C. Burnett, J.F. Holzrichter, and T.J. Gable, "Background speaker noise removal using combined EM sensor/acoustic signal signals," presented at the 138th Meeting of the Acoustical Society of America, Columbus, Ohio, Nov. 1999.

[6] R. Hu and D. V. Anderson, "Single acoustic channel speech enhancement based on glottal correlation using non-acoustic sensors," in *International Conference on Spoken Language Processing*, Jeju, Korea, Oct. 2004.

[7] D. Messing, *Noise Suppression with Non-Air-Acoustic Sensors*, Masters Thesis, MIT, Sept. 2003.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York.

[9] G.A. Darbellay and Igor Vajda, "Estimation of the information by an adaptive partition of the observation space," *IEEE Transaction on Information Theory*, vol. 45, pp. 1315–1321, May 1999.

[10] V.N. Vapnik, *Statistical Learning Theory*, New York: Wiley, 1998.

[11] Aravind Ganapathiraju, J. E. Hammake, and Joseph Picone, "Signal modeling techniques in speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 81, pp. 1215–1247, Sept. 1993.