The Use of Low-Frequency Ultrasound for Voice Activity Detection

Ian Vince McLoughlin

Abstract

An active detection system is developed which uses low-power low-frequency ultrasonic reflection to determine the lip state (i.e. whether open, closed or in between) of a human speaker and hence the presence of vocal activity. In operation, a small loudspeaker or sounder, located within a few centimetres of the lips, produces an excitation signal which is emitted towards the lips. A co-located microphone receives the signal reflected from the lip region. Even simple analysis of the reflected information reveals whether the mouth is open or closed. Given an excitation located above the normal frequency range of human speech, the method is unaffected by speech energy. if the excitation frequency is moved above the normal threshold of human hearing (i.e. an ultrasonic excitation), the method is inaudible. Careful placement of the excitation signal at the extreme low end of the ultrasonic range, allows its generation and analysis to be done with inexpensive off-the-shelf audio hardware.

This paper describes the techniques used, presents experimental details regarding the signals, then implement and evaluates a simple voice activity detector based on the technique.

Index Terms: Lip state detection, mouth state, low frequency ultrasound, voice activity detection

1. Introduction

A few published examples exist of speech systems making use of low-frequency (LF) ultrasound, mainly for speech augmentation. In one of the pioneering works, Tosaya and Sliwa used an ultrasonic signal injected into the vocal tract, from both outside and within the mouth [1], to improve the noise robustness of voice recognition. Soon after, Lahr used 28–100 kHz ultrasound along with video (and sound) in a trimodal recognition system [2]. In his system, ultrasound was inserted and detected from in front of the mouth, with the signal down converted to audio and appended to the feature vector obtained from audible speech. Douglass also used ultrasound as an ASR input channel in 2006 [3], including a similar excitation point. More recently, the current author proposed using a low frequency ultrasonic wideband chirp excitation, injected from in front of the face, to perform mouth state detection [4]. The technique was shown to be promising in terms of correctly distinguishing mouth open or closed (or partly open) states. However it was only evaluated for single vowels sustained over several seconds, since the chirp repetition rate could not exceed 2 Hz (due to hardware linearity limitations)[5]. In this paper, the generation and analysis techniques are extended to support a repetition rate of 10 Hz, and as a consequence, the system is fast enough for natural speech. This paper thus evaluates, for the first time, the use of LF ultrasonic reflection for voice activity detection (VAD).

1.1. Motivation

The primary motivation for a VAD employing ultrasound is that it can be robust to audible noise. For high levels of corrupting acoustic noise, especially when the noise shares statistical properties with the signal of interest (e.g. speech corrupted by multi-speaker babble), there may simply be insufficient information contained within a single audio signal to accurately detect the presence of individual speech. At lower noise levels, VAD is possible, but still challenging. In such cases, additional non-audible or non-vocal information may yield advantages in determining voice activity. In fact several non-audible VAD methods exist, including some that yield good performance, such as (i) using video of the lips [6, 7], (ii) Doppler based systems [8, 9] and (iii) Skin contact electrodes [10].

Video systems detect mouth opening and lip movement with good accuracy, however are unsuitable for use in the dark, or where a mouth is obscured, perhaps by beard or by sensor location. They also require a continuous video stream to be captured and analysed, which is significant in terms of power consumption and computational complexity. Doppler systems sense lip acceleration, also require careful placement and involve significant signal processing, however they do report excellent performance [8]. Unfortunately they require specialised sensors with narrow bandwidth. Physically, they yield only a single data point for mouth opening and mouth closing respectively, which may limit their accuracy. Skin contact electrodes also reportedly perform well [10], tend to have lower resolution than video but are more robust. Again, this requires specialised sensors, which must be placed appropriately to capture the data. Obviously they must touch the face to operate, and also require careful setup. Reportedly, they may not be suitable for all users [11]. Clearly each method suffers disadvantages, and no single one is suitable for all application scenarios.

This paper introduces a new approach that, like Doppler systems, also utilises ultrasonic information – although in this case it is wideband, and much lower in frequency. It is able to provide multiple data points even for single mouth open/close events, is of extremely low complexity and, as will be demonstrated, performs well even in the presence of wideband background noise.

1.2. Application

The idea of low-frequency ultrasonic reflection for lip state detection was originally introduced by the author in [4]. As a basic technique, it provides several important benefits including (i) by using inaudible signals, it can be imperceptible in use, (ii) the excitation is not only above the threshold of human hearing but above the range of normal human sound production. Thus it is robust to nearby speech, babble and similar noises. (iii) almost any audio device capable of sampling up to 48 kHz (or possibly...
even 44.1 kHz) can produce and analyse the signals, and analysis is not computationally taxing.

In noisy environments, robust VAD allows systems such as private mobile radio (PMR) and mobile phones to ignore voice-like sounds that are picked up during non-speech periods, and which would otherwise be processed as speech. Potentially, accurate lip-synchronous information could be used to improve speech processing, although this is not demonstrated in the current paper. Likewise, the technique shows potential for speaker validation: a speaker could be authenticated not only by the sound they produce, but also by the non-audible mouth movements, i.e. the patterns of mouth opening and closing. It would require that the subject is physically present during authentication, rather than simply a recording of the subjects’ speech. Again, this is not explored further in the current paper.

The emphasis of the current paper is to propose a novel VAD system which uses the ultrasonic reflection technique, and is highly robust to wideband background noise.

1.3. Contribution

This paper extends prior published work [4, 12] in three main ways: (i) a faster and broadband chirp excitation is proposed, which allows use with speech (previous systems could only work with single spoken vowels), (ii) adaptive spectral subtraction is used to self-calibrate the response of the system at the receiver, (iii) a new and simple frame-wise mouth state detection approach is developed, (iv) this ultrasonic mouth reflection technique is used for the first time to implement a VAD. (v) Experiments in acoustic background noise are reported in terms of accuracy and false triggering. Excellent performance will be demonstrated, even in high levels of noise.

1.4. Operation

A loudspeaker or sounder, placed a few cm from a user’s face, outputs a periodic wideband LF ultrasonic signal. This reflects from the face to be recorded by a nearby microphone. When the face, microphone and loudspeaker positions are fixed, the response of the received signal changes in accordance with the characteristics of the reflecting object – in this case the face. The major difference to the reflective state of the human face between a mouth open and a mouth closed condition is the mouth itself. When the mouth is open, incident ultrasonic energy couples into the resonant chamber formed by the mouth cavity and VT, and a large difference is evident in the reflected signal compared to when the mouth is closed. The physiology and physics of LF ultrasonic propagation into the mouth and VT have been presented in [4], and the human biological effects of the signals used have been discussed in [13, 14].

2. LF ultrasonic VAD

The LF ultrasonic signal source and detector are placed horizontally, located axially to the mouth, a few centimetres away. One major advantage of the system is that precise placement is unimportant (evaluated in [4]) although the transducers are ‘pointed’ towards the mouth. Fig. 1 shows the prototype hardware arrangement. The internal microphones of the Zoom H4n recorder are not used – only the small condenser microphone located slightly to the left of the chip source (which is a tweeter connected to a tube). The metal frame is simply to mount the transducers and protect them from damage.

A range of 14–21 kHz was used for the chirp in these experiments. These frequencies partially overlap the upper range of human hearing, especially for children, and are also partially within the low-frequency ultrasonic region. Apart from being readily generated by equipment with a sample rate of 44.1 kHz, this signal range was convenient for experimentation in that it remains partially audible and yet beyond the influence of significant vocally-produced speech energy. The chirp is a linear swept-frequency cosine pulse train of duration \( t = 200 \text{ ms} \) with each chirp set to linearly sweep the frequency range of \( f_1 = 14 \text{ kHz} \) to \( f_2 = 21 \text{ kHz} \) with constant amplitude:

\[
x(t) = \cos(2\pi(f_1 t + (f_2 - f_1)t^2/2\tau))
\]

Since the chirp is linear and flat, the amplitude envelope of the reflected signal will be proportional to its spectral response. Thus, timed from the start of the chirp, the amplitude at \( t = 0 \) gives the \( f_1 \) Hz response. That at \( t = \tau \) gives the \( f_2 \) Hz response and the \( (f_1 + f_2)/2 \) Hz response at \( t = \tau/2 \). This allows for low complexity frequency domain analysis without requiring the computational requirement of an FFT to be performed for each analysis frame.

2.1. Analysis

Fig. 2 shows a very simple differential time-frequency mesh plot of a single mouth close-open-close event (namely, speaking the vowel /ə/) recorded at the microphone. The mouth begins closed at \( t = 0 \text{ s} \), opens to approximately 1.5 s, achieves a full opening around 3.8 s and then closes again. It is shut again by 5.2 s. Fig. 2 actually plots \( r(t) - r_c(t) \), defined below in eqn. (2), for each 200 ms analysis frame.

![Figure 1: An illustration of the prototype hardware.](image)
During operation the raw microphone signal \( m(t) \) (sampled at \( F_s = 96 \text{kHz} \)) is bandpass filtered with a 43-order Butterworth IIR (passbands of \( f_1 \) to \( f_2 \), cutoffs 500Hz beyond those limits, 1dB passband ripple and 40dB attenuation). The resulting signal is demodulated to baseband, \( r(t) = m(t).\sin(2\pi f_c t) \) and down sampled to 32kHz. The mean response during the first and last \( N = 5 \) frames (which are known \textit{a priori} to be closed in this example), is computed as a reference,

\[
r_C(t) = \frac{1}{2N} \sum_{n=1}^{N} [r_{[n]}(t) + r_{[\text{end}+1-n]}(t)]
\]

Each chirp-sized frame of \( r(t) \) is then further reduced in dimension. In this paper, Welch’s power spectral method is used with 512 sample overlapping frames and a 1024 point FFT to obtain a smooth lower dimensionality \((L_{PS} = 1024/6)\) estimate of the frame shape, denoted \( W(f) \) – which is then to be classified as either ‘open’ or ‘closed’. In fact the author has experimented with classifying the raw input frame (with poor results due to the influence of many local imperfections and AWGN effects on the received signal), as well as several other methods of obtaining \( W(f) \). These include FFT, Hilbert transform, 3rd order Savitzky-Golay polynomial smoothing filter (non-causal but works well) and simple envelope detector.

\( W(f) \) is compared to the expectation of the closed-lip response vector, \( W_C(f) \), in order to make a decision regarding frame status (open or closed). \( W_C(f) \) is updated after two or more successive frames are classified as ‘closed’. The difference between open and closed frames is manifested by significant negative-going peaks in the response. These are accumulated to obtain a frame-by-frame scalar signature, \( S \). The negative-going peaks occur because opening the mouth causes significant resonances (zeros) to be introduced:

\[
S = \sum_{i=0}^{L_{PS}} \min \{W(i) - W_C(i), 0\}
\]

Decision \( D(k) \) for frame \( k \), is the thresholded frame-wise signature difference, \( D(k) = \{|S(k) - S(k-1)| > \gamma\} \) where \( \gamma = 2 \times \text{mean}\{S(k)\} \) for the results reported in this paper. A more sophisticated decision could have been used, however the performance of this simple method is good enough to demonstrate the potential of the VAD.

2.2. Performance

A volunteer was placed so that their mouth was located around 3 cm in front of the microphone and sounder, and asked to speak a series of TIMIT sentences in the presence of background noise. The sounder was set to output the chirp at a low amplitude, with the microphone and recorder (Zoom H4n, Zoom Corp., Tokyo, Japan) recording in 16-bits at 96 kHz sample rate. The recorder thus captured both the reflected ultrasonic signal, plus the speech of the volunteer. The speech part of the recording is lost during the down converting and anti-aliasing filter operations, so that only the reflected chirp signal remains for analysis. Fig. 3 (top) shows the 96 kHz time domain waveform plotted for several sentences, with the detection metric \( S \) and decision \( D \) shown below. Obviously, the decision output is NOT uniformly positive for every speech frame because the speaker naturally has to close their mouth occasionally when speaking. However, the quiet period between recordings does not exhibit any false detections (the volunteer was asked to breathe nasally).

Another volunteer was asked to speak and mime a sequence of words. This is shown in Fig. 4 where a Savitzky-Golay decision metric is plotted at the top along with frame-by-frame decision, the acoustic waveform is plotted in the middle and a 48 kHz bandwidth spectrogram at the bottom. The \textit{spoken} words are the first one and last two, whereas the \textit{mimed} words are the second and third. The absence of obvious vocal energy for the mimed words in clear from the waveform plot, which also reveals the repetitive chirps. These are also visible in the spectrogram which also reveals the spectral energy from the spoken words (but not the mimed ones). The implication is that the technique does not rely upon vocal energy for detection – and that the “voice activity” which is detected may be unvoiced, whispered or mimed. It is thus suitable for use in a Silent Speech Interface (SSI), as well as providing orthogonal information to a traditional VAD. To clarify, traditional VAD detects energy (such as formant sub-band energy) and is thus good for voiced phonemes, particularly vowels, which tend to have higher energy [15]. By contrast, the present technique is...
equally able to detect unvoiced phonemes, or whispers. It could replace a traditional VAD, or operate alongside it to improve final detection accuracy.

### 3. Evaluation

The analysis above has established that vocal energy is unnecessary for speech detection. However the converse question remains: does additive noise degrade the accuracy? An experiment was thus constructed to answer this question. Three normally speaking female (F) and four male (M) subjects were asked to recite TIMIT sentences, seated in front of the test apparatus in a soundproofed room. The methodology used was similar to the previous experiments described above. In this case, the speakers were reciting the sentences in the presence of wideband background noise at three SNR levels (-10 dB, 0 dB and +10 dB at the source). The volunteer speakers were asked to remain stationary, but were otherwise not restrained, and were asked to breathe nasally if possible.

The noise types used and test conditions were based on those in [8] (a Doppler ultrasound VAD). The lowest mean speech-to-noise SNR at the recorder was 2.6 dB, while the highest was 25.6 dB. The chirp-to-noise ratio varied from -6.2 dB to 0.2 dB.

Mean VAD accuracy over all of the experimental test material is shown in Table 1 for each of the tested noise types (after [8]). False detections are also reported, defined as any frame, in the gap between utterances, falsely identified as speech. Frames that are detected as ‘mouth closed’ during a speech period are compared to their neighbours. Three consecutive ‘mouth closed’ frames are required to flag the start of a non-speech period (which is then timed from the first of the consecutive ‘mouth closed’ frames).

From these tests, the mean accuracy in +10dB noise was 96.7%, for 0dB noise was 92.7% and 93.0% in -10dB noise. False detections were 1.5, 1.6 and 2.0% respectively.

Overall detection accuracy exceeds 90%, which is comparable to [8] (which used a very different technique, but similar experimental evaluation). The results do not depend strongly upon the presence of background noise.

### 4. Conclusion

This paper has proposed, tested and evaluated a novel VAD based on an LF ultrasonic excitation reflected from the mouth and lip area of a speakers’ face. The system is inherently silent to human hearing and is highly robust to audible background noise. The physical basis of the signals have been shown to very clearly distinguish between mouth open and mouth closed states. Also new in this paper is a low complexity adaptive VAD metric that performs well: accuracy exceeds 90%, even in the presence of speech-like babble noise. Further details of this work are available from www.lintech.org/savad

### 5. Acknowledgements

The theoretical framework for LF ultrasound was established by the author’s research team in Nanyang Technological University (NTU), Singapore. Thanks are due in particular to past PhD student Farzaneh Ahmadi and the many student volunteers in USTC and NTU who contributed to this work. The author is also grateful to USTC for the support of this research under grant no. KY210066002.
6. References


