Circuits, Systems, and Signal Processing, April 2015
(volume 34, issue 4, pp. 1279-1304)

Mouth State Detection From Low-Frequency Ultrasonic Reflection

Ian Vince McLoughlin · Yan Song*

Received: Jul. 2014 / Accepted: Oct. 2014

Abstract This paper develops, simulates and experimentally evaluates a novel method based on non-contact low frequency ultrasound which can determine, from airborne reflection, whether the lips of a subject are open or closed. The method is capable of accurately distinguishing between open and closed lip states through the use of a low complexity detection algorithm, and is highly robust to interfering audible noise. A novel voice activity detector is implemented and evaluated using the proposed method and shown to detect voice activity with high accuracy, even in the presence of high levels of background noise. The lip state detector is evaluated at a number of angles of incidence to the mouth and under various conditions of background noise.

The underlying mouth state detection technique relies upon an inaudible low frequency ultrasonic excitation, generated in front of the face of a user, either reflecting back from their face as a simple echo in the closed mouth state or resonating inside the open mouth and vocal tract, affecting the spectral response of the reflected wave when the mouth is open. The difference between echo and resonance behaviours is used as the basis for automated lip opening detection, which implies determining whether the mouth is open or closed at the lips. Apart from this, potential applications include use in voice generation prosthesis for speech impaired patients, or as a hands-free control for electrolarynx and similar rehabilitation devices. It is also applicable to silent speech interfaces and may have use for speech authentication.

Keywords Lip state detection · low frequency ultrasound · mouth state detection · speech activity detection · voice activity detection

National Engineering Laboratory of Speech and Language Information Processing,
The University of Science & Technology of China,
Hefei, Anhui 230027, China
*Corresponding author E-mail: songy@ustc.edu.cn

1 Introduction

The low frequency (LF) ultrasound (20–100kHz) range encompasses a diverse set of industrial and medical applications. Industrial ultrasound applications, which typically involve high power signals, mainly occupy this low frequency range. Apart from industrial uses, various therapeutic medical treatments make use of this range [5]. The few examples of low frequency ultrasound applied to human speech systems are mainly for speech augmentation. For example Tosaya and Sliwa [32] patented a system which applied ultrasonic signal injection to the vocal tract, both outside and within the mouth, to make the task of voice recognition more robust to audible noise. Meanwhile, Lahr [20] obtained an ultrasonic output from the vocal tract (VT) as the third mode of a trimodal voice recognition system, operating alongside audible voice and video of the lips, tongue and teeth. He used a position in front of the mouth as an injection point for a 28–100 kHz excitation, with the signal penetrating the VT through the mouth and teeth opening. The ultrasonic reflected output of his system was demodulated to the audible range and used directly as an input channel to a voice recognition system. Douglass [11], meanwhile patented a system using ultrasonic excitation to add value in improving the reliability of automatic speech recognition (ASR) and for speech transmission. His excitation points included in front of and inside of the mouth. Between them, these implementations pioneered the possibility of using LF ultrasound in speech processing. Inspired by these works, a physical framework was proposed for LF ultrasound propagation through the VT by the first author in 2010 [6].

The present paper introduces a very different use for LF ultrasound. Firstly it extends the theoretical framework for LF ultrasonic analysis of the vocal tract, simulating the VT response, for realistic excitation, before developing a novel application for lip opening detection. This derives a binary determination of whether the lips are open or closed, and is subsequently applied to the task of voice activity detection (VAD), which itself is explored in [21].

When in use, a sounder and a microphone are held within a few centimetres of a human face, facing towards the mouth. The sounder outputs a periodic wideband LF ultrasonic signal, which propagates through air towards the mouth. The ultrasonic signal is reflected from the face and is then picked up by the microphone as shown in Fig. 1. Consider deployment of the system in a mobile telephone: if the loudspeaker end of the mobile telephone is held to the ear during normal speaking operation, then the ultrasonic sounder and speaker could be installed in the opposite (mouth) end of the device, facing towards the mouth. With this arrangement, if the reflection area includes the region of the mouth, a large difference in reflected signal will be evident when the lips are open, compared to when they are closed. This difference is largely due to the resonant chamber formed by the open mouth cavity and VT. In this paper, we will firstly discuss the physics and physiology of these effects, evaluate these resonances using simulations and measurements on fabricated VT models, and then propose and evaluate a low complexity algorithm able to distinguish automatically between the mouth open and closed states.

Existing methods for automatically detecting lip opening or mouth shape do exist, but tend to be costly, require specialised equipment and be computationally expensive. Previously published techniques include the following:

1. Video cameras [18,25] pointed at the lips, analysed to detect movement or shape. Obviously, these do not work well in low-light conditions, or where a mouth is obscured. Furthermore, they require a continuous video stream to be captured and analysed, which involves significant power consumption and computational complexity in addition to an appropriately placed camera.
2. Doppler based systems for lip movement detection [17,15], which tend to have lower resolution than video based systems. These are considered to be relatively robust and high performance, but are sensitive to positioning. Specialised sensors must be placed appropriately to capture the data.
3. Skin contact electrodes [7,16] can reportedly provide good performance. The main disadvantages are that they must physically touch or be mounted on the face. They reportedly do not work well for all users and suffer from several other disadvantages [10].

By contrast, the low-frequency ultrasound based system described in this paper provides several potentially important benefits, including: (a) The system, using inaudible ultrasonic frequencies, can be imperceptible to the user and others in the vicinity of the user, (b) LF ultrasonic frequencies are not only above the threshold of human hearing but also above the range of normal human sound production. Consequently, the signals are robust to audible noises such as background speech, (c) The hardware required to generate, output, capture and process these signals is not specialised: in fact many smartphones are capable of handling the signals, leading to a potentially very inexpensive system for lip opening detection, (d) The system requires very little power to operate, much less than involved in placing a voice call, since the signals are of low energy, and only low-complexity processing is required, (e) The ability to lip-synchronise a speech processing system could lead to improved intelligibility and noise robustness, for example, by being able to ignore voice-like sounds picked up during an extended period when lips are not moving.

The underlying motivation of the present work has several applications, the main immediate examples being medical in nature. Some examples are:

1. Voice regeneration in speech impaired patients, including enabling hands-free lip-synchronous control of an electrolarynx, as a speech therapy aid, and for speech

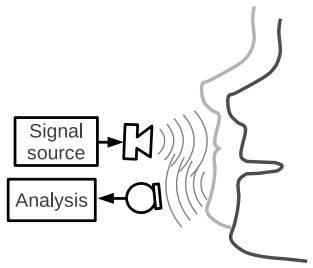

Fig. 1 Block diagram of basic mouth state detection system arrangement.

prosthesis design. The latter point includes systems able to detect the whisper-like voice of post-laryngectomised or trachaeotomised patients and convert this to more normal speech [26], since these systems suffer from the reduced signal-to-noise ratio and noise-like spectral characteristics of the detected voice, they are thus often unable to reliably determine speech onset and termination (especially in the presence of a noisy background).
2. Improved voice activity detection or speech activity detection (SAD) is also advantageous for mainstream speech communications systems that are required to operate in acoustically noisy environments. Virtually all speech communications systems, and many human-computer speech interface systems include a VAD which operates on the basis of audible acoustic input signals. Since continuous human phonetic production is related and synchronous to lip opening and closure, reliable knowledge of mouth movement has potential to enable more reliable VAD – a simple form of this is evaluated experimentally in the current paper (the VAD is developed further in [21]). The converse is also true: while lips remain closed, speech is unlikely. This point may also prove advantageous for power saving if complex and power-hungry speech analysis/transmission algorithms employed during a voice call can be paused, saving power, while a speakers' mouth remains closed but loud background noises occur which would otherwise trigger operation.
3. The technique also has potential for speaker validation. An authentication system may add non-audible mouth movement data to the recorded speech of a subject speaking into the system. The system would, in effect, not only authenticate based on the audible speech of a user but upon the patterns of mouth opening and closing. It would, for example, require that the subject is physically present during authentication, rather than simply a recording of the subjects' speech.

This paper investigates the physical basis of LF ultrasonic response within the human VT and mouth in Section 2, before analysing physical and computational models of the system in Section 3. This understanding is used to design a system able to automatically determine the lip opening state or mouth movement of a subject in Section 4, which continues to report on five experiments that were conducted to explore the novel detection method. In particular, a system is implemented to perform basic VAD, and then evaluated in the presence of acoustic background noise. Section 5 concludes the paper, and considers limitations of the technique and present study.

2 Analysis of low frequency ultrasonic propagation in the VT

For audible wave propagation, the VT is generally considered to be a linear system, which considerably simplifies the mathematics of describing the process. One question that arises is whether the linearity assumption is also valid for LF ultrasound propagating in the VT. If it can be verified as being a linear system, LF ultrasonic lip status detection – and other applications using sounds of this frequency range – can benefit from using a source-filter model to describe the excitation of the VT with an external ultrasonic source.

In order for LF ultrasonic propagation in the the VT to be linear, both the transmission medium and the structure of the system need to preserve the linearity criteria. These are thus considered in the following subsections.

2.1 Propagation through the vocal tract

Attributes of the medium include amplitude linearity, attenuation and dispersion effects. Although the amplitude response of signals propagated through exhaled air is known to be non-linear at high sound pressure levels, a system suitable for continuous human use avoids such levels, and thus maintains linearity due to only small amplitudes being involved [3]. The two remaining frequency dependant attributes of LF ultrasound propagation (in the air outside or inside the VT) are attenuation and dispersion effects. Attenuation has two components, namely absorption and scattering. While scattering of sound can be considered negligible in this application, absorption of ultrasound in air is considerably frequency dependant, increasing rapidly with frequency. However, limiting the frequency range of the current application to ultrasonic frequencies of less than 100 kHz, the attenuation coefficient does not exceed 3dB/m (ISO 9613-1, 1993) which can be ignored for the dimensions of the acoustic system in question (i.e. the typical VT length, surveyed extensively by [33]). The other precondition of linearity for ultrasound propagation in air is that the air medium should be an ideal gas in which the speed of sound is independent of sound frequency. For the frequency range of interest and the dimensions of the VT, the dispersive effects of the air medium bounded inside the vocal tract is negligible, despite their being higher concentrations of CO_2 in exhaled air (CO_2 is known to be dispersive [36])

Considering negligible effects of dissipation and dispersion, LF ultrasonic propagation in the VT can be approximated as a linear lossless process with any excitation source being similar in principle to excitation of the VT at the lips as in VT spectrometry [12,34] or as in the precise measurement of the VT impulse response in the audible domain [28,19]. On the other hand, the VT is a resonant cavity at LF ultrasonic frequencies that has one major difference with the audible case: at audible speech frequencies, due to the relatively large wavelengths involved, sound mainly propagates axially along the VT in one dimension only. As the frequency increases above a certain limit, cross-modes appear. For a simplified circular tube model of the VT, this limit is defined as $0.5861c/2a$ [13] where c is the speed of sound and a is the radius of the tube cross-section. For a relatively large VT of $15cm^2$ cross-section [33], it is around 4.7 kHz [29]. Consequently at higher frequencies, in addition to axial standing waves, resonant cross-modes appear. The high frequency sound then propagates in three dimensions with a much more complex resonance pattern [3].

2.2 The air-tissue interface

In addition to propagation through air, a system driving the VT with a LF ultrasonic excitation will involve the signals encountering one or more air/tissue interfaces. This is another potential source of non-linearity or signal attenuation. However, the reflection coefficient for an air-tissue interface at these frequencies is high,

about 0.99 due to a large impedance mismatch between the two mediums: the acoustic impedance of air ($Z_1 = 0.0004 \times 10^6$ Rayls) is much smaller than that of soft tissue ($Z_2 = 1.71 \times 10^6$ Rayls for muscle)[1]. So assuming negligible dispersive effects [35], airborne ultrasound almost completely reflects back from an air/tissue or tissue/air interface, and is constrained very well within the VT. Thus when the mouth is closed, the signal reflects very well from the face. When the mouth is open, the signal propagates into the VT, where it is largely constrained by the VT walls.

3 Modelling LF ultrasonic VT propagation

Section 2 established that LF ultrasonic propagation inside the VT can be approximated as a linear system. A mathematical formulation for this has been described previously by the first author [3].

The mathematical model derives from general wave equations of propagation with the use of a series of assumptions and simplifications. Firstly is that of irrotational flow in propagation through the VT. Although the assumption is challenged by rotational flows being present in turbulent air, and jet flows being present in the production of unvoiced utterances during speech, the non-linearity is commonly included as an attribute of the glottal source [27] in classical source-filter models of the human vocal system. In the LF ultrasonic case, the excitation source is not actually glottal in nature, and typically does not involve turbulent air jets. Therefore it is only when an LF ultrasonic signal happens to excite the VT simultaneously with the jet flow of an unvoiced utterance that the non-linearity would potentially occur, and in this case it does not immediately affect the signal injection point (which is located by the lips), instead affecting a much more distant reflection point of the system (i.e. the glottal region). It therefore requires a more minor approximation than that which is conventionally applied for audible speech systems.

For a linear model, we also assume only small disturbances in pressure and density as mentioned, which preserve the linearity of the medium – in addition to discarding effects of dispersion. Low power ultrasound is also much preferred over high power signals for system implementation in terms of potential bio-effects [5]. In fact, the LF ultrasonic equations of propagation through the VT are very similar to those for audible propagation [14], with the main difference being that the sound wave propagates in three dimensions, whereas the audible case is commonly considered to be mainly axial [6].

3.1 Modelling VT propagation

To further understand the LF ultrasonic propagation through the VT, finite element simulations were performed on models of the human VT. Since the standard 44-tube Kelly-Lochbaum model is known to be valid for sampling frequencies up to 60 kHz [8], this was used as the basis of the simulated VT structures. In total, six English

[1] The speed of sound is approximated to 1600 m/s in muscle and 343 m/s in air.

vowels were modelled[2]. The area functions and the length of the tubes were provided by [30], with the impedance of the VT walls made equal to the mean acoustic impedance of soft tissue in the human body (1.7×10^6 Rayl). Wave equations valid for LF ultrasound [2], were used to drive the propagation.

In order to verify the model structures as well as the basic modelling technique, the system was first explored in the audible domain where known resonance data, i.e. 'ground truth' exists. Resonances were computed from the finite element models, and compared with the simulated and natural resonances of the real VTs as reported by [30]. The results of this comparison are reported in [4]. The degree of correspondence was calculated between the simulated 44-tube model resonant frequencies, and the natural and simulated resonances. The resonances were compared since they relate to the important formant frequencies in voiced speech [22]. Results showed a very close correspondence for most resonances (apart from the case where lip opening is extremely small). This previous work therefore validated the basic technique in the audible domain, where comparative data is available, and is not reproduced here. In the present paper, the same technique will now be extended slightly upwards in frequency to the LF ultrasonic domain.

3.2 Modelling ultrasonic VT propagation

With the aim of modelling the lip opening detection system developed in this paper, an ultrasonic excitation was applied to the same VT models from a point in front of the mouth (the glottis was closed during all tests, as a perfect reflector, and the lips open into a free field). Several measuring positions, located in front of the lips at a distance of a few cm, separated laterally by distances of greater than one wavelength, were identified and the ultrasonic frequency responses evaluated at those locations. Fig. 2 plots the results of a time harmonic analysis of the 44-tube model resonances of one particular vowel /ɔ/ from 14–21 kHz in two locations. It is very clear that the resonant peaks match well at the measurement positions (and this is true for all measured positions), although the zeros do not necessarily align. The LPC envelope (which, as an all-pole representation is insensitive to the zeros in the response) matches for all measurement positions. Table 1 lists the resonance frequencies for each of six English vowels across the same LF ultrasonic frequency range, showing that the responses in each case are significantly different.

As a further step of verification, the six 44-tube vowel models were fabricated in epoxy using a 3D printer. The final VT models were excited by an ultrasonic loudspeaker in an anechoic chamber (a full description of the complete hardware setup will be given in Section 4). Resonance peak frequencies were determined for this physical setup. Resonances were also measured at audible frequencies (not reported here, but given in [1]). The physically measured resonances peaks were verified to have no more than 0.02% deviation from the simulation results of Table 1.

At this point, a physical basis has been developed for VT propagation, verified in the audible domain against finite element models of known VT shapes with published

[2] The six vowel geometries are: /i/, /æ/, /u/, /ɛ/, /ɔ/, /o/ as in heed, had, who, head, paw, and hoe respectively.

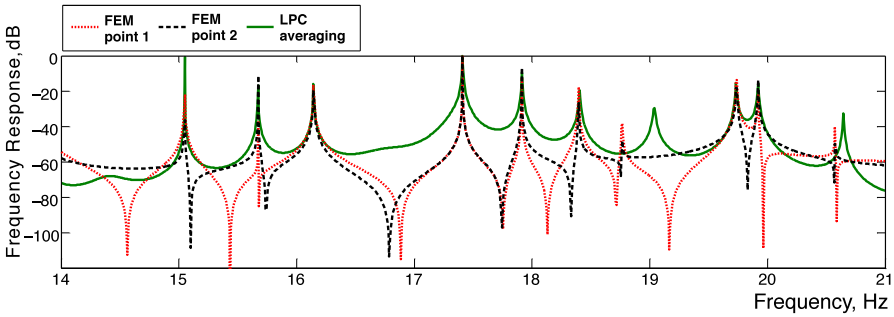

Fig. 2 Resonances from LF ultrasonic propagation through a 44-tube finite element VT model of vowel /ɔ/ (with excitation at the lips and the glottis considered closed), data from [1].

resonance data, and analysed in the LF ultrasonic domain. Three dimensional models were also built using the same VT shapes and acoustically tested at both audible and LF ultrasonic frequencies, to match both the published data and the finite element model responses. With confidence in the basic theoretical foundation and analytical techniques, subsequent sections will develop and then evaluate the lip opening detection system.

3.3 Simulation of open and closed mouth response

Before performing experimental and human subject testing, the system was simulated in its entirety using MATLAB. In this case, the simulation modelled the simple block diagram of Fig. 1, where a test subject began with mouth closed for 4 seconds, opened his mouth (and attached VT) in the configuration of the vowel /ɔ/ for 8 s, and then closed his mouth again for 4 s. A wideband Gaussian noise source signal was incident on his face, from which a signal reflected back to be captured by a microphone. Both

Table 1 Resonance frequencies of the 44 tube models for LF ultrasonic excitation at the lips (glottis closed), measured as described in [1].

/ɔ/	/æ/	/ɛ/	/i/	/u/	/o/
14,323	14,832	14,117	14,717	14,034	14,166
15,284	15,801	15,454	15,663	15,062	15,075
15,581	16,660	16,411	16,272	15,663	16,255
15,884	17,598	17,600	17,062	16,306	16,770
16,508	17,911	18,886	18,006	16,844	17,513
17,417	18,638	19,992	18,231	17,603	19,033
17,633	19,103		18,977	18,288	20,482
18,448	19,656		19,469	19,238	
19,372	20,248		20,337	19,429	
19,943				19,613	
20,227				20,273	

transducers were assumed to be perfect signal converters, and the air-tissue interface (see Section 2.2) assumed to be a perfect reflector. In the open mouth state, the VT of the simulated subject was in the configuration of the vowel /ɔ/ obtained from equal interval area functions of a human subject provided in [30]. The area functions were then bounded by a closed glottis and open lips, converted to reflection coefficients and then to LPC coefficients[3].

Given fixed area functions, the resulting frequency response would obviously be stationary during each phase. Therefore the simulation applied a random 12.5% fluctuation (zero mean, normal distribution) on each area function element every 0.5 s to model a plausibly realistic rate and degree of movement of the VT articulators. During the mouth opening phase, 50% of the incident signal was coupled into the VT, with the remainder being reflected by the face.

Fig. 3 shows a time-frequency domain visualisation of the simulation result for the frequency range 14–21 kHz. The range and plotting method are chosen to match the experimental output from human subjects presented later in Section 4, Fig. 6. Briefly, a stack of 160 short-time power spectra are plotted, derived from 10 non-overlapping analysis frames per second, over a duration of 16 seconds. The first and last few analysis frames represent the closed mouth response, whereas those in the centre represent the open mouth response. To generate this plot, a 'closed' spectral template was computed as the mean spectral response of the first 20 and the last 20 frames (known *a priori* to be closed). This was subtracted from the spectra of all analysis frames before computing the spectral power. The plot thus shows, during the open mouth period, a large and obvious power spectral difference compared to when the mouth is closed.

4 LF ultrasonic mouth status detection system

The proposed system comprises a source and detector, connected to signal generation and recording equipment respectively, which are to be located in front of the mouth of a human test subject, as shown in Fig. 1. In this section, the method is evaluated by constructing and analysing an experimental implementation.

In general terms, both source and detector are placed to act horizontally, located approximately axially to the mouth at a distance of a few centimetres. It should be noted that it is a highly advantageous aspect of the system that precise placement is not required (as will be discussed later). For the purposes of calibration, the test subject was replaced by a reflective board located where the mouth would be. When analysing the three dimensional models described in Section 3, these were placed with the 'lip' end around 5 cm from the source, the axis of the tube aligned horizontally with the source such that the closed 'glottis' end is then around 20 cm away. Fig. 4 shows the arrangement of the acoustic hardware system for a human test subject. In use, the source generates a LF ultrasonic excitation signal which is output towards the test subject and then reflected back from the mouth and face, to be captured by the detector.

[3] The conversion was made using the `lpcaa2rf()` and `lpcrf2ar()` functions from the excellent Voicebox package [9]

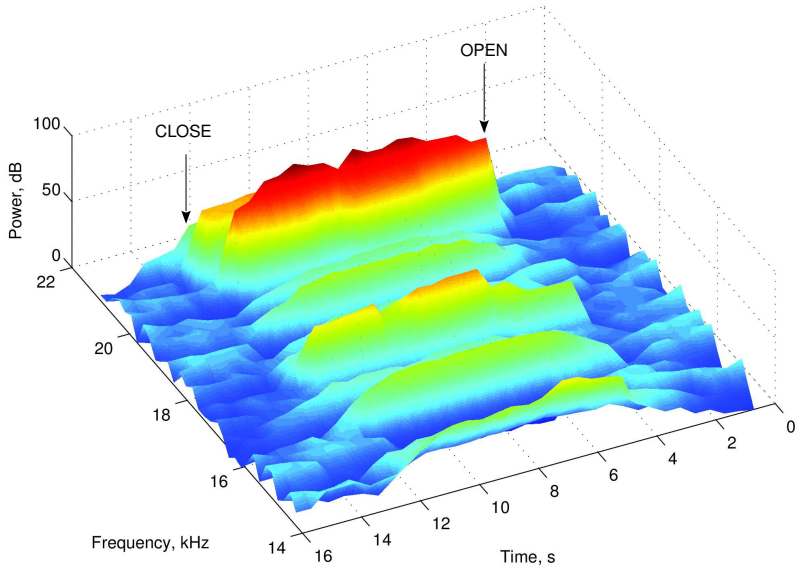

Fig. 3 Time-frequency domain representation of the closed/open/closed lip response for a simulated subject with VT in the configuration of the vowel /ɔ/ during the open mouth period. Peaks due to the resonances of the VT in the open mouth state are clearly evident in the response.

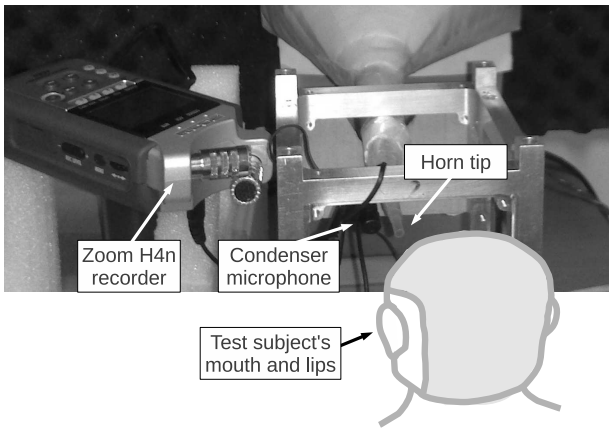

Fig. 4 Acoustic hardware of mouth status detection system (as deployed for Experiment 4). A schematic of the human head is used to illustrate the location of a test subject.

For the experiments in this paper, a relatively low frequency ultrasonic range of between 14 and 21 kHz was used, with some exploration up to 24 kHz. Although this range partially overlaps the upper frequencies of human hearing, especially for children, it is a convenient range for experimentation. Firstly, despite being partially

sub-ultrasonic, it remains outside the normal range of typical human speech and is thus unaffected by background speech (tested later), yet it allows a slight audible indication of the measurement signal, which was a significant help in the experimental procedure. Although any frequency below about 100 kHz could have been used (as discussed in Section 3), a lower frequency excitation signal can be generated by readily available audio hardware, and handled with many standard loudspeakers and microphones.

In this case, the excitation signal is a linear swept-frequency cosine (chirp) pulse train [23] with each chirp set to linearly sweep the frequency range of 14–21 kHz with constant amplitude. Using a linear chirp, the amplitude envelope of the reflected signal naturally varies in time according to its spectral response.

The hypothesis is that when the lips of the test subject are open in a vowel configuration, the VT appears acoustically in parallel with the free field baffled by the face, and that this is readily distinguishable from the closed lip state through an analysis of the reflected LF ultrasonic chirp. Furthermore that if this technique is to be usable for many of the applications mentioned previously, it should be relatively robust in terms of placement accuracy and be usable for the detection of speaking. A number of experiments are now described that have been conducted to explore and evaluate this novel technique.

4.1 Experiment 1 – lips open/lips close detection

A continuous sequence of 0.5 s long LF ultrasonic chirps were generated by a signal generator unit (DT Translations 9836, Data Translations Inc., Marlboro, Mass, USA) connected to and controlled by a laptop computer. The analogue output from the generator unit drove a horizontally mounted ultrasonic loudspeaker (Avisoft Scanspeak, Germany). This was coupled to an inverse horn acting as an ideal acoustic current source, located axially in front of the mouth of a test subject. The inverse horn was constructed, based on the description in [12], specifically to introduce no additional resonances to the flat frequency response generated by the loudspeaker. The generated signal was adjusted so that ultrasonic chirps recorded by the microphone at the tip of the horn had a relatively flat frequency response. Consequently, the speaker and the horn can be considered as an acoustic current source which is essentially independent of the load. For calibration purposes, a hard reflective plate was located in place of the subject's face, to ensure that the closed mouth response remained as flat as possible without making the system user-specific (i.e. the system hardware does not need to be adjusted for different test subjects). Once the initial system calibration was complete, the source signal and arrangement remained unchanged for the series of measurements. The procedure was conducted in an anechoic chamber with the test subject seated, facing the horn. An ophthalmic examination chin rest was used to maintain a horn-lip distance of about 5 cm.

The LF ultrasonic excitation enters the open mouth or is reflected back from the face in the closed lip state, and was captured and recorded using two wideband microphones (Zoom H4n, Zoom Corp., Tokyo, Japan, on board directional condenser microphones). The signals were sampled at 96 kHz with 16-bit precision.

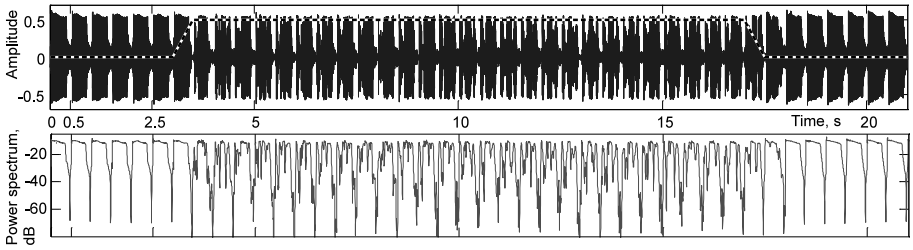

Fig. 5 (Top) time domain reflected waveform for a subject with lips closed, then open, then closed in a mimed vowel configuration. The dotted line represents a lip detection result. (Bottom) power spectrum of the reflected chirp, showing the increased 'peakiness' when the lips are open.

The system design would not be complete without considering standing wave resonances that may occur in the lip closed state between the closed mouth/face and the transducer. Although such resonances were found to be rare in practice, since the wavelength of sound in the frequency range of 14–21 kHz varies from 1.6 to 2.4 cm (the speed of sound being 343 m/s in the air at STP), they could potentially occur between the horn tip and the skin/lips. To suppress these resonances if they were to occur, a sound absorbing material was used to cover the horn body apart from the tip (not shown in Fig. 4).

4.1.1 Frequency response of closed and open lips

When the subject is in place, and opens his/her mouth, the resonances of the mouth cavity and vocal tract appear to cause strong variations in the frequency response of the measured spectrum. Since the excitation signal is a linear chirp, this is evident from the time domain envelope of the reflected chirp signal. By contrast, with the system set up as described, then when the mouth is closed, a relatively flat reflection of the ultrasonic excitation reaches the recorder without major resonance effects. This can be seen in Fig. 5 where the time domain reflected signal is plotted (top waveform), along with a simple power spectral representation (bottom waveform). During this test, the subject was instructed to began with their lips closed, open their mouth after about 3 s to form an /ɔ/ vowel shape, and then closed their lips again 15 s later. In this instance they mimed the vowel rather than spoke it aloud.

A dotted line is overlaid on the plot to indicate the open lip period (it is actually the output of an automated detection algorithm that will be described later). The plot very clearly demonstrates an obvious difference in reflected chirp between the two states of lips closed and lips open. Note that the periodic chirp signals at either end of the waveform plot are very similar in shape to the generated chirp measured at the tip of the horn (not shown).

To explore further, Fig. 6 plots a time-frequency representation of a single {closed–/ɔ/–closed} lip sequence where a male test subject began with closed mouth, then opened his mouth to mime the VT configuration of the vowel /ɔ/ for about 12 s before closing his mouth again. It can again be observed that the closed and open states have clear distinction, and the open lip state is characterised by the appearance of strong

resonances, as already predicted by the LF ultrasonic VT simulations. This plot, obtained from experimental data, appears very similar to that predicted by simulation in Section 3.3 and shown in Fig. 3 , although slightly more time-domain variation in spectral peaks is evident. Both figures were plotted using the same MATLAB signal analysis and plotting code, simply using different input data from simulation and experiment respectively.

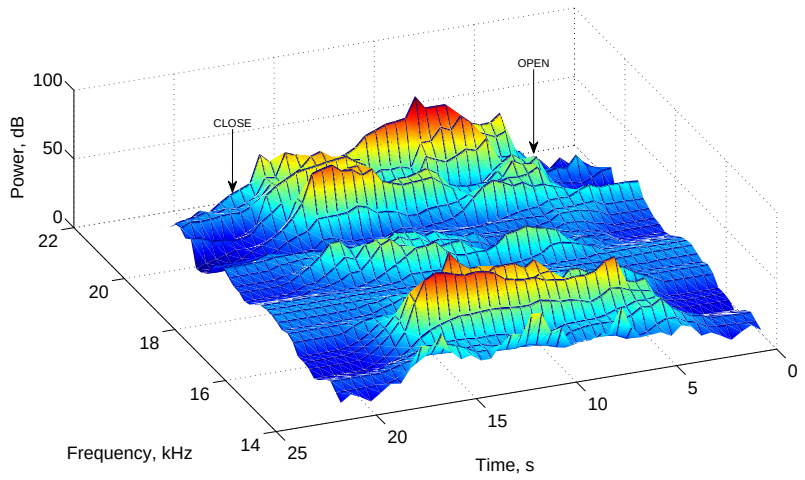

Fig. 6 Time-frequency representation of the closed/open /ɔ/ vowel/closed lip states. The arrows point to the locations of the lip opening and closing events.

4.2 Experiment 2 – lip state detection

In order to evaluate the effectiveness of lip state detection using this method, results were gathered from one female (F) and one male (M) adult subjects, both with unimpaired speaking ability and neither with facial hair. Signals were recorded for the LF ultrasonic chirp excitation being applied to a variety of mouth and lip opening conditions, under the same conditions as Experiment 1.

The lip state detection effectiveness was initially tested for a mimed configuration of 6 English sustained vowels. Three states were tested for each vowel, namely full opening of the lips and mouth, half opening (with lips partially open but no teeth visible from the front) and a semi-closed state (with lips mostly closed but visible teeth). It is acknowledged that the latter two states would naturally cause the vowel, if spoken, to change somewhat from the given prototype form (however in this experiment the vowel was not spoken or otherwise identified during the test so the actual sound involved is no more than hypothetical, and not a critical aspect of the experiment). In total, 60 recordings were captured for each user under each condition, making

a total of 360 recorded vowels for the two subjects. To ensure the performance of the system was independent of specific vowel utterances, the same experimental arrangement was then used to record results from the two test subjects asked to open and close their lips in a random manner (i.e. not in any specific pre-determined vowel configuration). 60 further samples were recorded in this way per subject.

4.2.1 Detection algorithm

An algorithm was implemented to detect the status of the lips from the above test recordings. The 14–21 kHz reflected chirp signal sampled at 96 kHz was first band-pass filtered and then demodulated to the baseband to cover a frequency range of 0–7 kHz, resampled at 14 kHz in the process. Next, the signal was segmented into 50% overlapping frames of length 0.5 s (given the 2 Hz chirp rate used for these tests). The beginning of each chirp pulse window was detected using autocorrelation between the known linear swept-frequency cosine signal and the reflected sound segment being analysed. Very strong autocorrelation peaks identified the beginning of each chirp. Since the linear source-filter theory applies to the LF ultrasonic and sub-ultrasonic excitations of the VT, this acts as a frequency-selective filter on the chirp signal (the source), when the mouth is open. Consequently, the time domain envelope of the reflected pulses become proportional to the frequency response of the VT cavity. This envelope, $P(f)$, is then extracted as the basis for detecting the status of the lips. The envelope, $P(f)$, was visible in the lower plot within Fig. 5.

Lip status was determined by the presence of greater number of peaks in the frequency spectrum in the open state. Since VT resonances appear in the open state response, the number of peaks increases significantly, and this attribute can therefore be used for decision making. Considering x_P and x_V to be the peak and valley values of $P(f)$ for each chirp pulse repetition respectively; μ_P and μ_V as the mean values of the peaks and valleys in that pulse respectively, a metric C is defined such that:

$$C = E\left[(x_P - \mu_P)^2\right] + E\left[(x_V - \mu_V)^2\right] \tag{1}$$

The result of this determination is plotted in Fig. 7 using the same raw data as that presented in Fig. 5. Peaks in the power spectrum are identified as small circles overlaid on the lower power spectral waveform, which then serve as the basis of the lip opening detection, with the result of simple threshold detector shown as a dotted line on the top waveform.

4.2.2 Lip state detection accuracy

The frame-wise binary lip state detection results for the {open/vowel/open} responses are listed for each of the tested vowel configurations in Table 2. Three lip states (full open, half open and semi closed) are identified separately for the male and female subjects. It is evident that extremely high accuracy is possible using this simple technique. There is also some variation of the results with the vowel used, indicating that it may be possible to use this technique to identify individual vowels or phonemes (in much the same way that lip-reading is performed based on lip shape). However

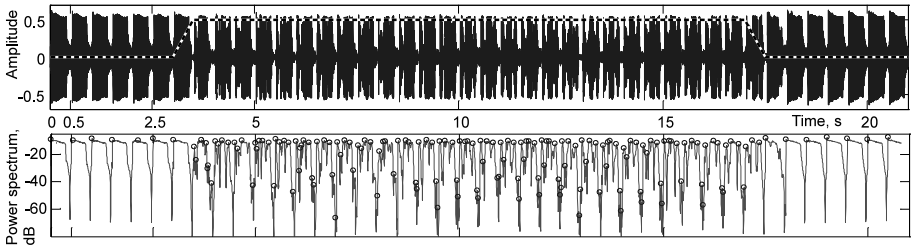

Fig. 7 As per Fig. 5, but indicating the detected spectral peaks and valleys in the reflected chirp signal power spectrum as small circles.

this has not been tested in the current paper, beyond noting that the lowest scoring results occur for the smallest lip openings (in a similar way that the smaller lip opening vowels resulted in the lowest correspondence between simulation and theory in Section 3.1). These results are consistent with the use of VT resonances as a detection feature – smaller lip openings naturally reduce the energy coupling into and out of the VT. A second test involved random opening and closing of the lips and mouth,

Table 2 Frame-wise percentage accuracy of the system for six English vowels with three degrees of lip opening, incorporating some results from [1].

Vowel	Subject	/ɔ/	/æ/	/ɛ/	/ɯ/	/i/	/o/
Full-	F	98.6	96.0	94.7	97.3	95.0	93.8
open	M	96.0	95.0	93.4	96.4	96.0	96.0
Half-	F	95.7	93.4	91.4	95.2	95.5	92.1
open	M	93.4	90.8	90.8	93.4	93.4	92.8
Semi-	F	96.0	93.4	88.2	90.8	82.0*	91.6
closed	M	94.4	90.8	85.6*	88.2	93.4	88.2

* The mouth opening was sometimes less than 3 mm wide open

where subjects were asked to have their lips either closed, or open their mouth in a sustained random shape (not necessarily in a vowel configuration) before closing it again. These results yielded a frame-wise accuracy exceeding 90% for the female and 94% for the male over the 60 instances per subject.

4.2.3 Experimental conditions

Several variations in experimental conditions were explored during the course of these tests. Firstly, the chirp signal amplification was adjusted across a range of around 40 dB_{SPL}. In general, it was found that decreasing the signal power tended to weaken the resonance response as less energy propagated into the mouth and VT. During the tests, the gain was adjusted such that sufficient signal energy was present to generate resonances in the VT, with the target gain range set to match typical speech amplitudes of the subjects. During the setup phase the LF ultrasonic signal was replaced or alternated with an equivalent 0–7 kHz audible chirp. The signal level was maintained at a comfortable amplitude experienced by the subjects. Increasing

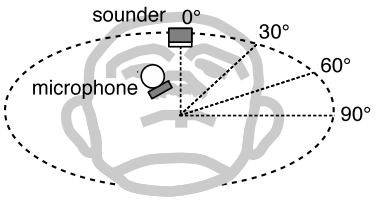

Fig. 8 Recording points located at different angles to the mouth axis in the horizontal plane, at $0°$, $30°$, $60°$ and $90°$.

the gain too much was found to cause adjacent resonances to merge together, however higher gain conditions were not explored, constrained by provisions in the human ethics approvals for this work. In practice, a clear and very wide window of usable gain was noted, and was easy to determine empirically. For all experiments described in the paper, it was neither necessary nor desirable to adjust gain for different subjects.

Precise placement of the source and detector was also unnecessary (this aspect will be explored later), and even the chin rest was discarded in later experiments as careful head orientation and horn tip–lip distance were not found to be critical to performance up to about 10 cm, as long as the tip of the horn remained clearly pointing at the mouth of the subject, and the positioning was pseudo-stationary during two chirp periods. Thus, in general terms, the system was found to be robust to alignment, speaker and microphone placement, signal level and background noise.

In order to assess this further, additional experiments were undertaken. The first is an evaluation of the effects of positioning accuracy on performance. Beyond that are experiments to determine the effect of background noise.

4.3 Experiment 3 – effect of angle of incidence

As predicted by finite element modelling (and shown in Fig. 2), the main sensitivity of this system is not its recording location. However the direction (angle) of the excitation point with respect to the mouth could still affect the performance of the system. Therefore, an experiment was constructed to verify the effect on performance of adjusting the angle of excitation in a horizontal plane around the head.

A subject was seated in front of the horn, using the chin rest, with a distance of just under 5 cm between the lips and the horn tip. The angle of the horn was varied with respect to the axis of the mouth and lips. Excitation points were located in a horizontal arc equidistant to the mouth, at angles of $0°$ (directly in front), then $30°$, $60°$ and $90°$ to the side as shown in Fig. 8. A total of 40 vowel and 40 random articulations were recorded at each excitation angle. Excitation points were adjusted repeatedly between articulations, and measured to be about +/- $10°$ around the desired angle.

The effect of increased angle of incidence was to reduce the influence of VT resonances on the open lip responses. This is very evident in Fig. 9 which shows time-frequency plots of one instance of the {closed-open-closed} sequence (random articulation) for the four different angles of excitation. The observation of the VT-

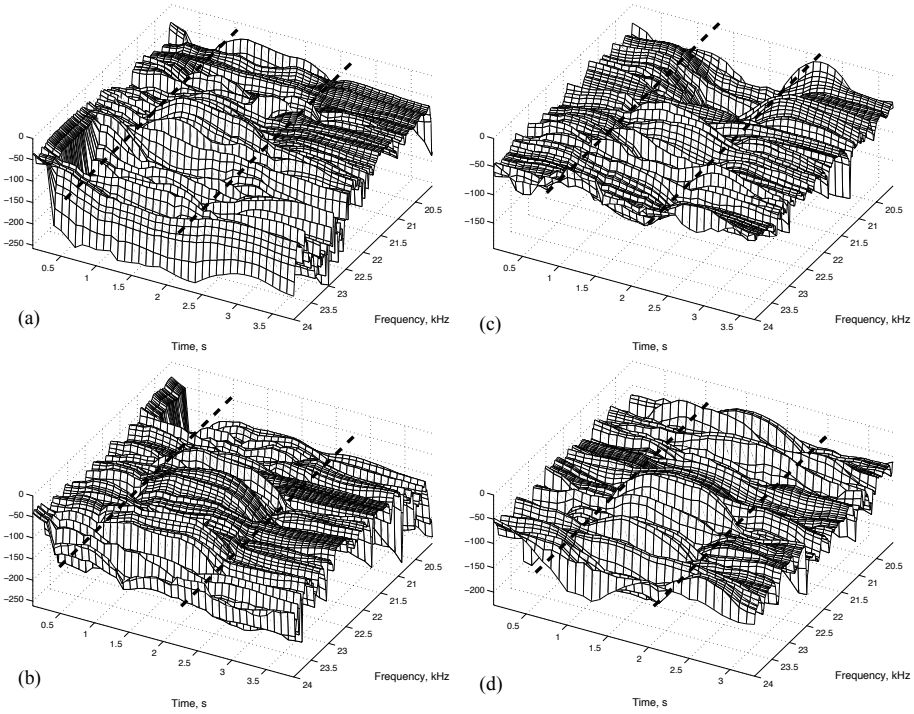

Fig. 9 Time-frequency plots of the {closed/open/closed lips} response recorded at side angles of a) 0°, b) 30°, c) 60° and d) 90°, with approximate transition instants marked by dotted lines.

induced resonances is most clear in the frontal excitation, but moving from frontal to 90°, the spectral resonances tend to reduce in amplitude, as well as becoming smoother. However, the two states (open/closed lips) remain easily distinguishable in each case.

The frame-wise binary lip open/lip close determination was repeated for the angle-of-incidence recordings, yielding accuracy figures reported in Table 3. While performance is obviously better when the source is axial with respect to the lips, and degrades as angle increases, even the perpendicular condition yields very good results exceeding 80% accuracy.

Table 3 Frame-wise percentage accuracy for different angles of exposure for vowel configuration and random openings. Further detail can be found in [1].

Angle	0°	30°	60°	90°
Vowels	93.6	92.9	90.4	86.7
Random	92.0	89.4	85.3	81.5

4.4 Experiment 4 – voiced and unvoiced speech

The experiments reported up to now involved constrained positioning of the test subject using an ophthalmic chin rest. They also required the test subject to carefully control their mouth and lip shape in a sequence of closed lips, mouth open in a vowel shape, then closed lips. Although the random mouth opening test was more realistic, the conditions were still relatively unnatural, and some of the equipment was quite specialised. Therefore two further experiments were designed: the first experiment considers the effect of voicing and non-vowel shapes, while the second assesses the technique in terms of its accuracy as a voice activity detector, and specifically evaluates the effect of background noise on performance.

In these tests, conducted in a soundproofed room, identical chirp signals were generated using MATLAB and replayed using a Zoom H2 recorder (Zoom Corp., Tokyo, Japan). An audio amplifier designed to work with sampling frequencies up to 192 kHz (Yamaha RX-V371, Yamaha Corp. Japan) was used to amplify this analogue output. The resulting signal was reproduced using the 19 mm high frequency tweeter unit from a Kef C3 loudspeaker, rated up to 40 kHz (GP Acoustics Ltd, Kent, UK), coupled to an inverse horn, similarly to the previous experiments.

Subjects was seated with their mouth axially located in front of the ultrasonic source horn, at a distance of between 1–6 cm, and the microphones moved slightly to the side of their face, at a distance of between 2–8 cm. Since exact positioning is not critical, the primary consideration was to ensure that both the open and closed lip states remained roughly axial with the horn tip. While movement was not physically constrained in any way, subjects were asked to remain as stationary as possible with respect to the horn tip.

Again, the ultrasonic excitation was placed to enter the mouth when the lips are open (or reflected back from the face in the closed lip state) and be captured using two wide band condenser microphones (Zoom H4n, Zoom Corp., Tokyo, Japan). The recorder was set to sample at 96 kHz with 16-bit precision.

Since the results up to now involved subjects mouthing vowels, rather than speaking them, it was considered necessary to investigate whether the lip state detection technique would also operate for spoken phonemes, and for non vowels. Thus an experiment was devised in which subjects were asked to speak a sequence of short words with at least a one second pause between each utterance. Some words were to be mouthed (i.e. not vocalised) rather than spoken. For this test, the chirp excitation repetition rate was increased from 2 Hz to 10 Hz. All users had unimpaired speaking ability, no facial hair and were aged between 20 and 45 years old.

Fig. 10 reproduces some of the signals from one test. The upper plot reproduces the time domain waveform of the reflected chirp signal. There are 56 full repetitions of the 14-21 kHz chirps shown in the waveform (each with relative amplitude of approximately 0.05). During the recording period shown in the figure, a male subject recites five letters of the alphabet beginning with ‘E’. Letters ‘F’ and ‘G’ are mimed rather than spoken. The lower plot shows a spectrogram of the recorded signal (sample rate is 96 kHz so the normalised frequency spans 0 Hz to 48 kHz), with the linear chirps easily visible as diagonal lines. The three spoken letters are also very evident as the vocal energy can be seen clearly in the spectrogram. By contrast, the spectrogram

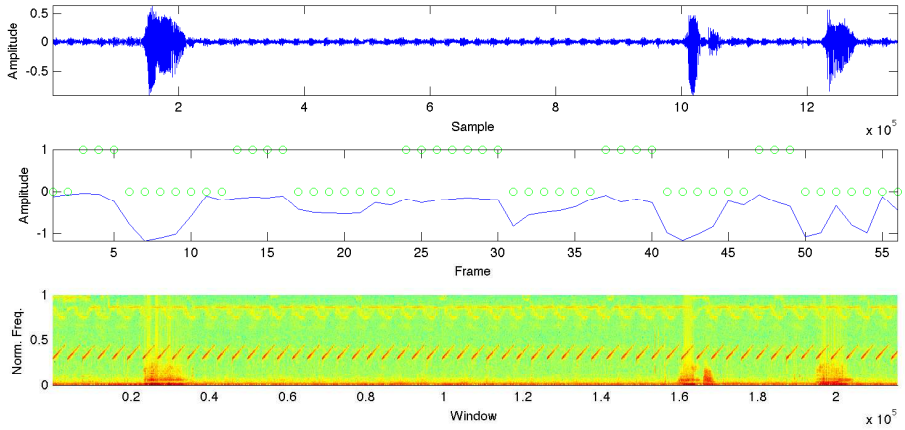

Fig. 10 (Top) time domain waveform of the reflected signal showing a mixture of low amplitude chirps and recorded speech (a male subject reciting the alphabet from 'E' to 'I' with 'F' and 'G' being silently mimed rather than spoken). (Middle) frame-wise detection metric with decision output plotted as circles, 0 indicating open lips and 1 indicating closed lips. (Bottom) 48 kHz spectrogram of the reflected signal, showing the chirps as diagonal lines, as well as revealing the clear spectral presence of the voiced letters.

shows very little evidence of the two mimed letters. In fact, evidence of the presence of the mimed letters, as well as the spoken letters, is carried in the diagonal reflected chirps. The solid line in the middle plot shows the summed difference between frame-by-frame Welch's power spectral density estimates taken of the down-converted and bandpass filtered chirps (i.e. processed as per the detection algorithm discussed previously). The circles indicate the output of the decision metric: zero indicating a lip open condition, and 1 indicating a lip closed condition. Apart from the two frames at the start, when the algorithm has not yet stabilised, the detection is perfect for both voiced and non-voiced phonemes. This result is typical: in all tests there was no instance where vocal information (i.e. audibly spoken words) was found to improve the detection accuracy.

4.5 Experiment 5 – vocal activity detection

Having established that vocal information does not improve detection accuracy for spoken or mouthed phonemes, the question that remains is whether the converse is true: does background noise degrade the accuracy of the system? Apart from determining robustness to background noise, we will also evaluate whether the method can be used as a VAD (note that a further VAD application is evaluated in [21]).

Strictly speaking, the role of a VAD is to detect the presence of voice, but some of the potential applications of this technology require the detection of whispered or mouthed speech – for example in speech prosthetics for voice-loss patients. Although we maintain the terminology, the precise task being tested will be to detect spoken,

whispered or mouthed sentences without using audible voicing information, i.e. using only the frequency response of the 14–21 kHz reflected chirp signal.

Conducted in a soundproofed room, this experiment used the equipment described above for Experiment 4. In addition to the previous arrangement, wideband background noise was played within the room while subjects were asked to read aloud a sequence of sentences. The test conditions were designed to match those used in the Doppler-based VAD evaluation by Kalgaonkar et. al. [17]. Wide band background noise recordings, consisting of Office, Car, Babble, Speech and Music backgrounds[4] were added acoustically (not mixed in digitally) to create a background far-field noise environment.

Three different noise SNRs were used, set to -10 dB, 0 dB and +10 dB at the source. Due in part to the Lombard effect [22], and the closeness of the recorder microphone to the mouth, the lowest mean speech-to-noise SNR measured at the recorder was 2.6 dB, while the highest was 25.6 dB. The chirp-to-noise ratio measured at the same place varied from -6.2 dB to 0.2 dB.

Three female (F) and four male (M) subjects (normal speakers, no facial hair), recorded individually, were asked to read aloud a sequence of sentences with a silent gap of 5 to 10 s between each. Sentences were selected from TIMIT [31], and presented in a random, balanced, sequence of 20 with multiple repetitions of the list to fill a fixed recording duration of 250 s per subject. During the test, subjects were asked to remain stationary, but were not physically constrained. They were requested to keep their lips closed when not speaking – the reason being that the presence of audible speech was used to judge lip opening 'ground truth', since no other information was recorded which could provide this information, apart from the chirp resonances under evaluation.

4.5.1 Vocal activity detection performance

Detection analysis was performed using the detection measure described previously. The expectation operator was automatically adjusted periodically during the analysis to adapt to varying levels and type of background noise. An example recording from this test is reproduced in Fig. 11, showing the time domain waveform at the top of the plot. For each chirp-sized analysis window, aligned at a chirp start using autocorrelation as discussed previously, a determination was made as to whether it occurred during a lip-open or lip-closed period. The absolute value of this detection metric is plotted in the lower window of Fig. 11. Simple thresholding of this metric is used to derive a hard lip open/lip closed decision (plotted as circles in the lower window). The first third of the recording corresponds to a +10 dB background noise period, the middle third is at 0 dB and the final third corresponds to -10 dB.

To examine further, the first 20% of the recording is reproduced in Fig. 12, comprising seven spoken TIMIT sentences interspersed with gaps during which time the

[4] Office and Car recordings were obtained as 96 kHz, 24-bit and 32-bit sample files from Freesound.org (nos. 108695 and 193780 respectively), recorded on Tascam DR-100 mk-II using on board directional condenser microphones (TEAC Corp., Tokyo, Japan). Other recordings were made by the author using the on board directional condenser microphones of a Zoom H4n (Zoom Corp., Tokyo, Japan), recorded at a 96 kHz sample rate with 16-bit resolution. The original recordings are available upon request.

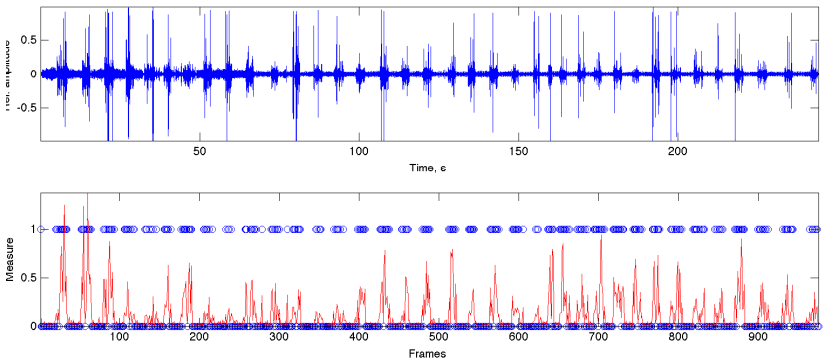

Fig. 11 (Top) time domain waveform of the reflected chirp recorded in background noise while a subject is reading a list of TIMIT sentences, separated by silent periods. (Bottom) the solid line plots the detection metric, with frame-wise decision indicated by circles. The background noise reduces in amplitude during the test.

lips were closed. For the first sentence shown (*"Just long enough to make you feel important"*), the mean speech-to-background noise ratio is 17 dB, but is -3.9 dB for the seventh sentence (*"I know I didn't meet her early enough"*), corrupted by high levels of background speech. Note that the decision metric (the lower plot) does not remain consistently high during the 'speech' periods – and naturally so because the lips are not consistently open during this entire period. For the speaker shown in this figure, all sentences were easily detected, with no false detections, irrespective of noise type and level.

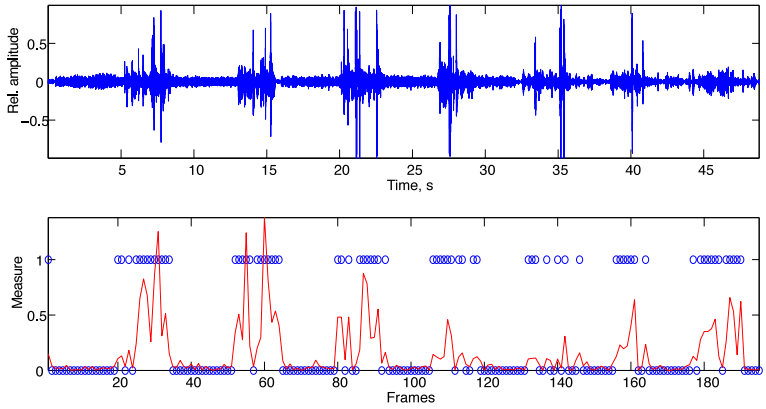

Fig. 12 An expanded version of the first seven spoken sentences from Fig. 9, more clearly showing the interrelationship between the detection metric and frame-by-frame hard decisions (bottom) to the time domain waveform (top).

A summary of results is presented in Table 4 in terms of the mean frame-wise VAD accuracy, for each of the noise levels and types, averaged over all speakers.

False detections are also reported. Clearly, the detection accuracy significantly exceeds 90% in most cases, and is comparable to the performance reported by [17], despite the fact that their work was completely different: it was a fixed-frequency 40 kHz Doppler-based mouth movement detector (we do not use Doppler analysis, or specialised sensors, but rely on a chirp at a much lower ultrasonic frequency range, depending on VT and mouth cavity resonances to distinguish lip state). There is no strong evidence from these results that performance depends upon either noise type or noise level, although it was noted that some test subjects scored consistently worse than others. In total, spurious detections occurred on average in 1.8% of frames during the test[5]. Many of these mis-detections were either isolated frames, or were extensions beyond a speech period – both of which would typically be ignored in a VAD system implementation [22]. On average, 3.8% of TIMIT sentence utterances were mis-detected, with a significant number attributed to a single test sitting by one subject (the figure reduces to 2.1% if that subject is excluded from the results).

Table 4 Mean frame-wise detection accuracy and false detection rate (i.e. spurious events, also reporting standard deviation), for voice activity detection of spoken TIMIT sentences in various levels of background noise.

Condition	Accuracy %	False % (σ)
Car	96.5	2.1 (1.2)
Music	94.2	1.1 (0.5)
Speech	94.1	1.6 (0.2)
Babble	91.7	1.8 (0.6)
Office	94.2	1.7 (1.0)
+10 dB noise	96.7	1.5 (0.7)
0 dB noise	92.7	1.6 (1.1)
-10 dB noise	93.0	2.0 (0.6)

4.5.2 Application consideration

The low frequency and low amplitude signals used in these techniques can be generated by standard audio equipment such as smartphones and even MP3 players (which can often sample at up to 48 kHz), however any technique with possible medical applications requires careful considerations of safety – particularly one which may involve long-term or continuous human exposure, such as in a voice prosthesis. Thus, the bio-effects of this range of low-frequency ultrasonic signals have been evaluated [5] and extended by proposing a modification of the standard mechanical index to apply to the low frequency ultrasonic region [24].

The signal generation and analysis algorithms involve frame-wise continuous processing, rather than batch processing, and are thus well-suited to real time implementation. Various detection metrics have been tested: the use of linear swept frequency chirp excitation means that computationally costly conversion to the frequency domain is unnecessary during operation.

[5] Note that, since the system detects lip opening rather than speaking, it is possible that some of these false detections did actually correspond to non-speech lip opening events if the subject opened their lips, for example to breathe through their mouth.

5 Conclusion

Low frequency ultrasonic excitation of the human vocal tract from in front of the face is a novel technique with attractive characteristics of being inherently silent to human hearing as well as being robust to audible noise. In general terms, the technique has the potential to be used as an additional modality in systems that regenerate voice for speech impaired patients, or for silent speech interface implementations. The low frequency ultrasonic signals used in this technique do not require specialised hardware – they can be generated, captured and processed by simple audio hardware such as that included in smart phones, and the processing itself is of relatively low complexity.

Propagation characteristics of low frequency ultrasound in the vocal tract were shown to provide a linear lossless system which can be described by the common source filter theory of (audible) speech production, but in a three dimensional propagation scenario. Finite element modelling of ultrasonic propagation inside the vocal tract predicted the occurrence of several resonances within the frequency range used. Three dimensional tube models were constructed for six English vowels and tested to validate the simulations: these models also experienced resonances, at very similar frequencies. In both cases, the resonances at low ultrasonic frequencies are denser than those at audible frequencies, in part due to the presence of transverse resonance modes at ultrasonic frequencies. This information was then used to describe a system able to detect the lip state of a human mouth, in particular whether the lips are open or are closed, by analysis of a chirp excitation signal reflected from the mouth region of the face. Results showed that the system is able to accurately determine lip state under a variety of conditions, as well as excitation angles (i.e. the angle between the signal generation device and the human mouth). Possible future applications of this system include hands-free control of electrolarynx and rehabilitation devices, silent speech interfaces and audio watermarking. The system was tested experimentally using both male and female test subjects, and shown to be capable of distinguishing mouth state with a high accuracy, using a simple and low-complexity detection algorithm. The lip state detection algorithm was then implemented as part of a voice activity detector, and evaluated in several levels and types of wide band background acoustic noise, while TIMIT sentences were spoken by male and female test subjects. Detection accuracy was shown to exceed 90%

This research is a first step in exploring the potential of low frequency ultrasonic analysis for speech-related systems such as silent speech interfaces, and primarily for speech impaired patients, with a particular emphasis on post-laryngectomised patients. Example data and MATLAB files for analysis may be downloaded from `http://www.lintech.org/savad`.

5.1 Limitations of the present study

The aim of this paper is to explore the potential of a new method of detecting lip-state using a novel excitation signal and method of analysis, and to extend this to a voice activity detection application. Results are generally positive, leading to the

conclusion that this technique is worthy of further study and development. Future studies should, in particular, address the following limitations:

- The lip state tests included vowels and vowel-shaped mouth openings, in which the presence of vocal tract resonances were detected to distinguish between lip-closed and lip-open states. Later tests included consonants, words and sentences, however the presence of vocal tract constrictions above the glottis as well as glottal closure itself were not evaluated in isolation.
- A larger variety of test subjects should be analysed since there is some evidence to suggest that certain users enjoy higher accuracy than others. Furthermore, none of the test subjects had facial hair (which could reasonably be expected to improve accuracy by further smoothing the closed-lip response in contrast to that for open lips and mouth), and none were from a target group of speech-impaired patients.
- The author was unable to identify individual vowels based on the precise resonance positions found in the chirp response. In fact, this is something noted previously [4]. Although it was not an aim of this paper, the ability to identify vowels (or preferably a range of individual phonemes) from their response would be extremely useful for speech recognition systems.

6 Acknowledgements

Some of the data for this paper was recorded and processed at the School of Computer Engineering, Nanyang Technological University (NTU), Singapore by student assistants Farzaneh Ahmadi, Mark Huan, and Chu Thanh Minh. Their contribution to this work is gratefully acknowledged, particularly the PhD research of Farzaneh Ahmadi [1]. Thanks are also due to Prof. Eng Siong Chng of NTU, and Jingjie Li of USTC for their assistance with the experimental work.

References

1. Ahmadi, F.: Voice replacement for the severely speech impaired through sub-ultrasonic excitation of the vocal tract. Ph.D. thesis, Nanyang Technological University (2013). URL `http://repository.ntu.edu.sg/handle/10356/52661`
2. Ahmadi, F., Ahmadi, M., McLoughlin, I.V.: Human mouth state detection using low frequency ultrasound. In: INTERSPEECH, pp. 1806–1810 (2013)
3. Ahmadi, F., McLoughlin, I.V.: The use of low frequency ultrasonics in speech processing. Itech Book Publishers (2009)
4. Ahmadi, F., McLoughlin, I.V.: Measuring resonances of the vocal tract using frequency sweeps at the lips. 2012 5th International Symposium on Communications Control and Signal Processing (ISCCSP) (2012)
5. Ahmadi, F., McLoughlin, I.V., Chauhan, S., ter Haar, G.: Bio-effects and safety of low-intensity, low-frequency ultrasonic exposure. Progress in Biophysics and Molecular Biology **108:3** (2012)
6. Ahmadi, F., McLoughlin, I.V., Sharifzadeh, H.R.: Autoregressive modelling for linear prediction of ultrasonic speech. In: INTERSPEECH, pp. 1616–1619 (2010)
7. Arjunan, S.P., Weghorn, H., Kumar, D.K., Yau, W.C.: Vowel recognition of English and German language using facial movement (SEMG) for speech control based HCI. In: Proceedings of the HCSNet workshop on Use of vision in human-computer interaction - Volume 56, VisHCI '06, pp. 13–18. Australian Computer Society, Inc. (2006)

8. Beautemps, D., Badin, P., Laboissihere, R.: Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data. Speech Communication **16**, 27–47 (1995)
9. Brookes, M., et al.: Voicebox: Speech processing toolbox for matlab. Software, available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html (1997)
10. Calhoun, G.L., McMillan, G.R.: Hands-free input devices for wearable computers. In: Proceedings of the Fourth Symposium on Human Interaction with Complex Systems, HICS '98, pp. 118–. IEEE Computer Society (1998)
11. Douglass, B.G.: Apparatus and method for detecting speech using acoustic signals outside the audible frequency range. United States Patent and Trademark Office: United States (2006)
12. Epps, J., Smith, J.R., Wolfe, J.: A novel instrument to measure acoustic resonances of the vocal tract during speech. Measurement Science and Technology **8**, 1112–1121 (1997)
13. Eriksson, L.J.: Higher order mode effects in circular ducts and expansion chambers. J. Acoustical Soc. America **68:2**, 545–550 (1980)
14. Fouque, J.P.: Wave Propagation and Time Reversal in Randomly Layered Media. Springer (2007)
15. Freitas, J., Teixeira, A., Dias, M.S.: Towards a silent speech interface for portuguese. Proc. Biosignals (2012)
16. Jorgensen, C., Dusan, S.: Speech interfaces based upon surface electromyography. Speech Communication **52**(4), 354–366 (2010)
17. Kalgaonkar, K., Hu, R., Raj, B.: Ultrasonic doppler sensor for voice activity detection. IEEE Signal Processing Letters **14**(10) (2007)
18. Kaucic, R., Dalton, B., Blake, A.: Real-time lip tracking for audio-visual speech recognition applications. In: B. Buxton, R. Cipolla (eds.) Computer Vision ECCV '96, *Lecture Notes in Computer Science*, vol. 1065, pp. 376–387. Springer Berlin / Heidelberg (1996)
19. Kob, M., Neuschaefer-Rube, C.: A method for measurement of the vocal tract impedance at the mouth. Medical Engineering & Physics **24**, 467–471 (2002)
20. Lahr, R.J.: Head-worn, trimodal device to increase transcription accuracy in a voice recognition system and to process unvocalized speech. United States Patent and Trademark Office: United States (2002)
21. McLoughlin, I.: Super-audible voice activity detection. Audio, Speech, and Language Processing, IEEE/ACM Transactions on **22**(9), 1424–1433 (2014). DOI 10.1109/TASLP.2014.2335055
22. McLoughlin, I.V.: Applied Speech and Audio Processing. Cambridge University Press (2009)
23. McLoughlin, I.V., Ahmadi, F.: Method and apparatus for determining mouth state using low frequency ultrasonics. UK Patent Office (pending) (2012)
24. McLoughlin, I.V., Ahmadi, F.: A new mechanical index for gauging the human bioeffects of low frequency ultrasound. In: Proc. IEEE Engineering in Medicine and Biology Conference, pp. 1964–1967 (2013)
25. Rivet, B., Girin, L., Jutten, C.: Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures. Audio, Speech, and Language Processing, IEEE Transactions on **15**(1), 96–108 (2007)
26. Sharifzadeh, H.R., McLoughlin, I.V., Ahmadi, F.: Speech rehabilitation methods for laryngectomised patients. In: S.I. Ao, L. Gelman (eds.) Electronic Engineering and Computing Technology, *Lecture Notes in Electrical Engineering*, vol. 60, pp. 597–607. Springer Netherlands (2010)
27. Sinder, D.J.: Speech synthesis using an aeroacoustic fricative model (PhD Thesis). The State University of New Jersey (1999)
28. Sondhi, M.M., Gopinath, B.: Determination of vocal-tract shape from impulse response at the lips. J. Acoustical Soc. America **49:6**, 1867–1873 (1971)
29. Story, B.H.: Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract (PhD Thesis). The University of Iowa (1995)
30. Story, B.H., Titze, I.R., Hoffman, E.A.: Vocal tract area functions from magnetic resonance imaging. J. Acoustical Soc. America **100:1** (1996)
31. Texas Instruments: TIMIT database (Texas Instruments and MIT). a CD-ROM database of phonetically classified recordings of sentences spoken by a number of different male and female speakers (1990)
32. Tosaya, C.A., Sliwa, J.W.: Signal Injection coupling into the human vocal tract for robust audible and inaudible voice recognition. United States Patent and Trademark Office: United States (1999)
33. Vorperian, H.K., Wang, S., Chung, M.K., Schimek, E.M., Durtschi, R.B., Kent, R.D., Ziegert, A.J., Gentry, L.R.: Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study. The Journal of the Acoustical Society of America **125**, 1666 (2009)

34. Wolfe, J., Garnier, M., Smith, J.: Vocal tract resonances in speech, singing and playing musical instruments. Human Frontier Science Program Journal **3**, 6–23 (2009)
35. Zangzebski, J.A.: Essentials of Ultrasound Physics. Mosby, Elsevier (1996)
36. Zuckerwar, A.J.: Speed of sound in fluids. In: M.J. Crocker (ed.) Handbook of Acoustics, pp. 61–71. Wiley-Interscience (1998)