Low Frequency Ultrasonic Voice Activity Detection using Convolutional Neural Networks

Ian McLoughlin[1,2], *Yan Song*[2]

[1]The University of Kent, School of Computer Science, Rochester, Kent, UK
[2]The University of Science and Technology of China, Hefei, Anhui, China

ivm@kent.ac.uk, songy@ustc.edu.cn

Abstract

Low frequency ultrasonic mouth state detection uses reflected audio chirps from the face in the region of the mouth to determine lip state, whether open, closed or partially open. The chirps are located in a frequency range just above the threshold of human hearing and are thus both inaudible as well as unaffected by interfering speech, yet can be produced and sensed using inexpensive equipment. To determine mouth open or closed state, and hence form a measure of voice activity detection, this recently invented technique relies upon the difference in the reflected chirp caused by resonances introduced by the open or partially open mouth cavity. Voice activity is then inferred from lip state through patterns of mouth movement, in a similar way to video-based lip-reading technologies. This paper introduces a new metric based on spectrogram features extracted from the reflected chirp, with a convolutional neural network classification back-end, that yields excellent performance without needing the periodic resetting of the template closed-mouth reflection required by the original technique.

Index Terms: Voice activity detection, speech activity detection, ultrasonic speech, SaVAD

1. Introduction

The recently published Super-audible voice activity detector (SaVAD) [1] used low frequency ultrasound (LFUS), in a 20Hz to 24kHz band, to sense mouth open or closed status of a subject from a facially-reflected chirp. The method was shown to yield discriminative signals from the two states at various angles of incidence and sensor distances from the face. The physical signals and method were explored further and simulated in [2] and [3] then developed into a voice activity detector (VAD) in [1].

The SaVAD equipment comprises a loudspeaker or sounder and a microphone. These are placed within a few centimetres of a human face, and could be mounted within the body of a mobile telephone used during a telephone conversation. Because the SaVAD signals are outside the human hearing and speech frequency range, they do not cause or suffer from interference when used during a voice call. In use, the loudspeaker, oriented towards the mouth region of the users' face, outputs a periodic LFUS chirp. This LF ultrasonic signal is then reflected from the surface of the face and received by a microphone mounted nearby, as shown in Fig. 1. When the face, microphone and loudspeaker are pseudo-stationary with respect to each other, standing waves are possible, primarily dependent upon the straight-line distances between source and receiver. If the reflection area includes the mouth, a significant difference in reflected signal is evident when the lips are open, compared to when they are closed. This main physical cause of the change in reflection characteristics is the resonant chamber formed by the mouth cavity and vocal tract which affects only the signal during the 'mouth open' condition.

A number of signal analysis techniques for SaVAD were explored in [2] while a simple signal processing-based detection metric was introduced in [1] and used to implement a VAD. The underlying LFUS signal propagation characteristics and linear predictive analysis methods were explored in [4].

1.1. Contribution

The present paper uses the same physical arrangements as the previously published systems, but develops a new set of features based on spectrogram images which are particularly suited to a convolutional neural network (CNN) based classifier. The CNN is trained on recorded segments containing both open and closed mouth conditions, and evaluated against both the original SaVAD technique [1] and the standard G.792 VAD in the presence of background multi-speaker babble. The CNN-based LFUS VAD is shown to significantly outperform both other methods, particularly in high noise conditions. In addition, the effectiveness of a CNN trained for a single user is compared to training for all individual users, and this highlights the ability of the CNN system to operate even when the physical alignment between user, microphone and loudspeaker changes – a particular weakness of the previous techniques.

1.2. Impact

The active LFUS chirp signal is inaudible to users, and is low power. It also has the great advantages of being compatible with many consumer-targeted digital audio hardware solutions, including those in modern smartphones, which can sample at up to 48 kHz in 16 bits. Similarly, it can be output by most standard micro-loudspeakers of reasonable quality and is detectable by

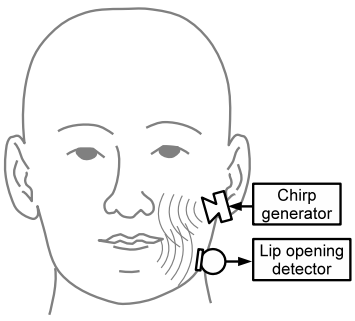

Figure 1: Block diagram of the SaVAD operating arrangement.

typical high fidelity microphones.

When implemented in a mobile telephone, this technology yields the ability to determine mouth movement patterns of the user. In the presence of high levels of background noise, current devices are unable to determine whether a sound picked up by the microphone originates from the user's speech, or from the background noise. The LFUS VAD signal provides a clear indicator of whether the user is currently speaking as well as a potential indication of syllabic rate, in terms of instances of voiced phoneme production.

The basic LFUS VAD also works for whispered or mimed speech – it does not require the presence of voicing – and thus can be a useful technique for silent speech systems.

2. SaVAD and the new spectrogram image feature

The SaVAD excitation signal is a linear cosine chirp of duration τ which spans a frequency range f_1 to f_2:

$$x(t) = cos\{2\pi(f_1 t + (f_2 - f_1)t^2/2\tau)\} \quad (1)$$

and is repeated continually with no pause during operation. For testing purposes, subjects are seated such that their lips are placed in front of the SaF signal source, at an axial distance of between 1–6 cm, and the microphones located slightly to the side of their face, at a distance of between 2–8 cm. The exact positioning of the microphone is less critical [5], and both angle and distance effects been investigated in [2, 1]. During recording, it is important to ensure that the SaF excitation is able to enter the mouth of a subject when their lips are open, or is reflected back from their face when lips are closed.

Microphone input signal, $m(t)$, is sampled at rate F_s (where $F_s \geq 2f_2$). This is first bandpass filtered using a 33-order infinte impulse response (IIR) filter matching the f_1 and f_2 extents, before being demodulated to a baseband frequency, $m'(t) = m(t).sin(2\pi f_1 t)$, and then down sampled. The frequencies and other system parameters used in practice are shown in Table 1.

The filtered and down-sampled received reflected signals are first cross-correlated with the transmitted prototype chirp to determine precise timings, before being divided into frames $s(n)$ of size $N = F_s' \times \tau$ samples. These blocks are then formed into a spectrogram $f(l,k)$ by stacking power spectra obtained from highly overlapped (by $Ws - Ol$ samples) Hamming windowed $w(n)$ regions of size Ws, using a P-point FFT. Spectral element k for the lth window is thus:

$$f(l,k) = \left| \sum_{n=0}^{P-1} s(n).w(n).e^{-j2\pi nk/P} \right| \quad (2)$$

Each spectrogram f is then correlated across both dimensions with its predecessor (i.e. the spectrogram from the previous chirp) f' to provide a measure of spectral change, yielding a correlation image, $c(l,k)$;

$$c(l,k) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m,n) f'(m-k, n-l) \quad (3)$$

for $(1-M) \leq k \leq (M-1)$ and $(1-N) \leq l \leq (N-1)$, given that $f(l,k)$ has a size of M by N. The resulting correlation matrix, $c(l,k)$ has size $M' = 2M+1$ by $N' = 2N+1$. This is then normalised and augmented by its marginals

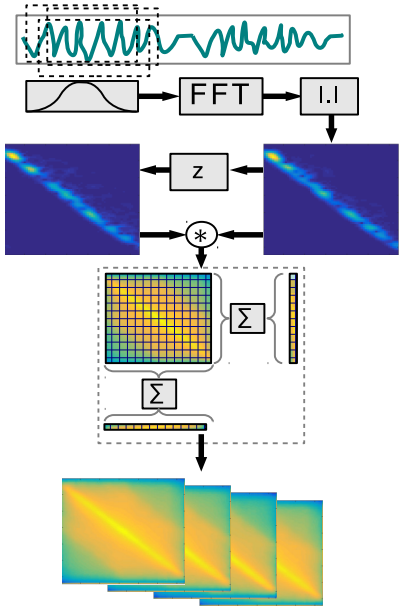

Figure 2: Feature extraction process showing initial spectrogram formation, differential correlation and formation of marginals.

(i.e. summed column vector and summed row vector) to form a $2M+2$ by $2N+2$ feature vector matrix, $v(l,k)$ as follows:

$$\begin{bmatrix} \zeta c(0,0) & \zeta c(0,1) & \dots & \zeta c(0,N') & \sum c(0,n) \\ \zeta c(1,0) & \zeta c(1,1) & \dots & \zeta c(1,N') & \sum c(1,n) \\ \dots & \dots & \dots & \dots & \dots \\ \zeta c(M',0) & \zeta c(M',1) & \dots & \zeta c(M',N') & \sum c(M',n) \\ \sum c(m,0) & \dots & \dots & \dots & \sum c \end{bmatrix}$$

where $\zeta = 1/max\{c(m,n)\}$. A diagram of the feature extraction process is given in Figure 2.

3. CNN structure

CNNs are multiple layered neural networks that consist of convolution, subsampling and fully interconnected layers. In general, convolution and then subsampling layers are alternated before a final fully interconnected layer forms the output into the final classification form. The network complexity (and hence capability) is relatively high due to the large amount of connectivity, although the use of shared weights within layers acts to reduce the number of parameters that need to be trained prior to operation. However, CNNs still generally require quite a large

Table 1: Specifications used in the SaVAD experiments.

$f_1 = 20$ kHz	$f_2 = 24$ kHz
$F_s = 96$ kHz	$F_s' = 8$ kHz
$\tau = 0.1$ s	$P = 127$
$Ws = 100$	$Ol = 11$
Output device:	KEF C3 tweeter
	(GP Acoustics, Kent, UK)
Input device:	Zoom H4n
	(Zoom Corp., Tokyo, Japan)
Test subjects:	4 female & 4 male
Noise levels:	-10, -5, 0, 5, 10dB babble

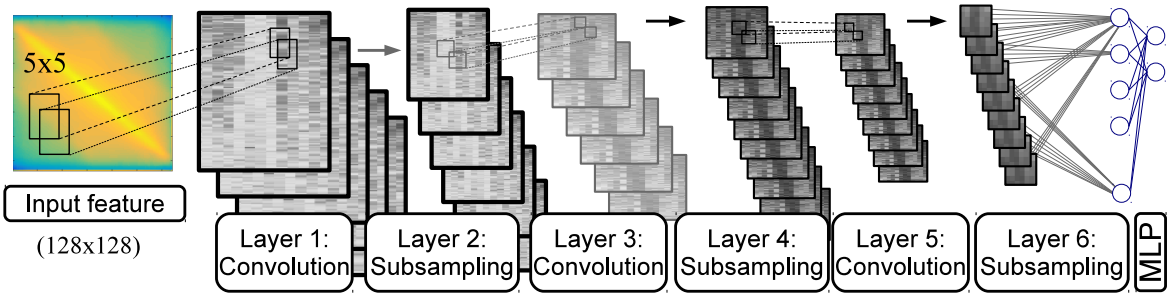

Figure 3: CNN structure used for classification of the spectrogram-based features.

amount of training data in comparison with some simpler machine learning algorithms. For a convolutional layer $l-1$, we form layer output maps from

$$\mathbf{x_j^l} = f\left\{\sum_{i\in M_j} \mathbf{x_i^{l-1}} * \mathbf{k_{ij}^l} + b_j^l\right\}, \tag{4}$$

where $\mathbf{x_i^{l-1}}$ is the ith input map, $\mathbf{x_j^l}$ is the jth output map, M_j represents a selection of input maps and $\mathbf{k_{ij}^l}$ denotes the kernel which is being used [6].

Subsampling layers are simpler than the convolutional layers, $\mathbf{x_j^l} = f(\beta_j^l \nabla(\mathbf{x_i^{l-1}}) + b_j^l)$ where $\nabla(.)$ represents the process of sub-sampling and both β and b are bias terms [6].

The final output layer is fully interconnected, effectively being a dual layer multi-layer perceptron (MLP) network. Its input layer size depends upon the total number of nodes in the final CNN subsampling layer, and its output size is defined by the number of classes. The CNN can be learned just like an MLP using gradient descent by following the back-propagation algorithm. Since units in the same feature map share the same parameters, the gradient of a shared weight is easily computed as the sum of the shared parameter gradients.

CNNs are widely applied in image processing [7, 8], with generally good results. They have also been applied successfully to ASR and other speech fields [9, 10].

A spectrogram is an image containing different patterns relating to the time-frequency distribution of the analysed sounds. When differentially correlated, the image of a reflected chirp highlights the areas of similarity (and hence dissimilarity) between the spectrograms of neighbouring reflected features. Although the underlying chirp signal appears as a diagonal high energy line in a spectrogram, the effects of the resonant cavity are to introduce zeros along an axial stripe centred on that diagonal line, in potentially any location or distribution. Thus local relationships are important, but these have only weak absolute locality, which is commonly known to be a strength of CNNs.

The final structure of the CNN used for SaVAD classification is shown diagrammatically in Fig. 3, with the major features listed in Table 2. The CNN toolbox [11] is used for the experimental implementation of the classifier.

Table 2: Specification of the CNN.

layer 1, convolution with 6 output maps, kernel size=5
layer 2, subsampling by 2
layer 3, convolution with 3 output maps, kernel size=5
layer 4, subsampling by 2
layer 5, convolution with 2 output maps, kernel size=3
layer 6, pass through
layer 7, fully connected with two output classes
batchsize=50, epochs=4

4. Experiments and results

4.1. Evaluation material

The initial recorded material used for this evacuation is a subset of the test material used for the simple VAD in [1], namely recordings from four female and four male volunteers. These recordings were made in a soundproofed room in the presence of added wideband background noise, with volunteers asked to read a repeating sequence of 20 TIMIT sentences with a mean gap of approximately 7 seconds gap between sentences. Unlike in the previous system, the data is partitioned into subsets for training and testing, totalling 50.3 and 12.6 minutes of speech for each subset respectively. During speech, the SaVAD equipment was positioned to output a chirp, and record the reflected response, recorded simultaneously with a voice-band recording of the subjects' speech. Subjects were seated and asked to minimise their head movements if possible, remaining within about 5 cm of the SaVAD loudspeaker cone and microphone.

Ground truth was obtained by a manual analysis of the voice-band recording, tracking sound power in the range 0-4 kHz, and thus serves as a measure of speaking rather than a measure of mouth opening. This is framed in alignment with the chirps, and summed to yield two classes of 'speech' and 'no speech', with intermediate states split between the two classes.

Training (and testing) data is prepared, one feature per chirp, in accordance with the process described in Section 2 and used to train the CNN along with the ground-truth. Only four epochs were necessary to achieve discriminative results, with a batchsize of 50. Given 10 chirps per second, the total training data comprised more than 30,000 features, with each feature being the 2D correlation between the current chirp spectrogram and previous chirp spectrogram, augmented with marginal information (sum-of-rows and sum-of-columns).

During evaluation, additional noise was added at levels of between -10 and 10 dB SNR, using the babble recording from [1] and following the methodology of that experiment. To compute SNR, the 'signal' refers to the mean power level of the 0 to 4 kHz original recording (i.e. mainly speech plus additional acoustic noise), while the 'noise' power refers to the mean power level of the 0 to 4 kHz noise recording. Due to the speech-like nature of the noise, the LFUS region noise power is at least 18dB below the low frequency region power. To be clear, the entire recording of wideband babble noise is added, concatenated if necessary to match the test recording length, at

the given levels, corrupting both the voice band recording and the LFUS region used for SaVAD.

4.2. Evaluation method

The evaluation compares the results from three methods. First is the G.729 VAD, referred to here as **G.729**, using the standard G.729C+ Appendix II VAD hangover, framed into 0.1 s analysis frames to match the evaluation unit of the SaVAD system. This is applied to the voice band 0-4 kHz recordings that were used to obtain the ground truth, although with the different levels of additive noise, and compared against the ground truth. The second method, referred to here as **Ultra**, is the basic signal processing approach to SaVAD classification used in [1], applied to the LFUS band recordings with added noise, and compared against the ground truth. The third method, referred to here as **CNN**, uses the same noisy data as Ultra, but reforms the classification using a trained CNN. For both the original Ultra and the new CNN methods, the raw binary mouth open/closed classification output is smoothed using a Savitzky-Golay smoothing filter [12] with a 3rd order polynomial over a 41-sample window and 31-sample window respectively and then thresholded to transform it to a simple binary estimate of voice activity.

Performance was evaluated using the standard criteria of Beritelli etc. al. [13], counting four types of error as follows:

- *Front-end clipping* (FEC) which counts errors made when transitioning from a non-speech region to a speech region.
- *Mid-speech clipping* (MSC) is the proportion of speech frames erroneously classified as being non-speech.
- *OVER* counts errors made when transitioning from speech to non-speech.
- *Noise detected as speech* (NDS) is the proportion of non-speech frames which are misclassified as being speech.

These mutually exclusive counters were incremented for each erroneous frame, then divided by the total number of frames in the test. Their sum yields the total error rate for each condition.

4.3. Results and discussion

For the first experiment, a CNN is trained separately for each recording session and used to classify voice activity from the spectrogram-based features. The results, obtained for additional SNR levels ranging from -10 dB to +10 dB, are shown in Fig. 4.

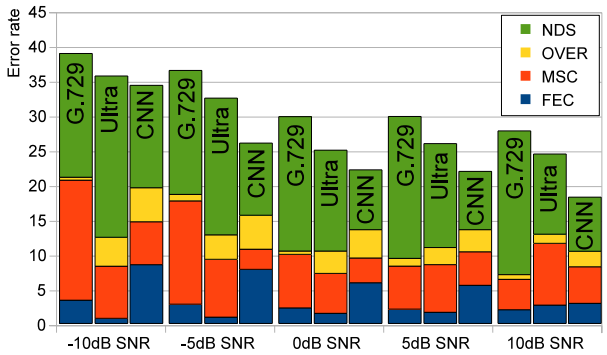

Figure 4: Stacked bar chart showing the component error rates for three VAD methods in the presence of additional SNR.

Table 3: Overall error rate for mismatched training conditions.

	-10dB	-5dB	0dB	5dB	10dB
G.729	41	36	32	30	29
Ultra	37	33	30	26	25
CNN[1]	31	31	29	26	20
CNN[2]	35	33	32	27	19

[1] Including test material corresponding to the training data.
[2] Excluding test material corresponding to the training data.

It is clear that the CNN method proposed in this paper significantly outperforms both the previous SaVAD based method (Ultra) as well as the G.729 VAD for all noise conditions. Interestingly, it has a relatively higher FEC and OVER score than other measures - indicating a delay in detecting the start or end of a speech region, which may be due to the causal nature of the spectrogram formation (i.e. forming a spectrogram from past analysis windows) or to the differential nature of the measure. Note that the error types between Ultra, CNN and G.729 are quite different: For example G.729, has a low rate of OVER errors, even in high noise levels, Ultra contributes few FEC errors, while CNN contributes the lowest rate of NDS errors. This evidence suggests that a metric combining all techniques, may outperform each individual metric.

A further experiment was conducted where the CNN was trained using material from only a single recording (i.e. one sitting, one user) but used to evaluate other recordings. In this case, the training material is obviously not indicative of all conditions, and thus CNN performance is reduced significantly. Table 3 reports average results where the CNN is trained using frames from one recording session, but evaluated using other session. Separate CNN results are shown for whether the remaining unused frames from the training session are included in the evaluation score or not. Performance is shown to degrade significantly compared to the results in Fig. 4 (although is still generally better than the alternative methods), particularly when the unused test session frames are not included in the evaluation. These findings indicates that the trained CNN includes some user-specific information, which leads to the interesting possibility that the SaVAD information maybe useful for speaker verification.

5. Conclusion

The paper has proposed the use of a convolutional neural network (CNN) for voice activity detection in the low-frequency ultrasound region using reflected chirps from the mouth region of the face. It has been evaluated for recorded speech in the presence of background noise and compared with the popular G.729 VAD operating on audible information, as well as a signal-processing based VAD published previously which makes use of the same low-frequency ultrasonic information as the proposed classifier. Results show that it outperforms both G.729 as well as the previous VAD metric under all tested conditions where the system is trained with user-specific information, but degrades when training and testing are mismatched. The proposed CNN-based classification technique is shown to be extremely robust to environmental noise, particularly that associated with speech, such as babble,

6. Acknowledgements

The authors acknowledge the support of National Natural Science Foundation of China (grant 61172158) and Chinese Universities Scientific Fund (grant WK2100060008) for this work.

7. References

[1] IV. McLoughlin, "Super-audible voice activity detection," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 9, pp. 1424–1433, Sept 2014.

[2] IanVince McLoughlin and Yan Song, "Mouth state detection from low-frequency ultrasonic reflection," *Circuits, Systems, and Signal Processing*, pp. 1–26, 2014.

[3] Farzaneh Ahmadi, Mousa Ahmadi, and Ian Vince McLoughlin, "Human mouth state detection using low frequency ultrasound," in *INTERSPEECH*, 2013, pp. 1806–1810.

[4] F Ahmadi, IV McLoughlin, and HR Sharifzadeh, "Linear predictive analysis for ultrasonic speech," *Electronics letters*, vol. 46, no. 6, pp. 387–388, 2010.

[5] Farzaneh Ahmadi, Ian Vince McLoughlin, and Hamid R Sharifzadeh, "Autoregressive modelling for linear prediction of ultrasonic speech," in *INTERSPEECH*, 2010, pp. 1616–1619.

[6] Jake Bouvrie, "Notes on convolutional neural networks," 2006.

[7] Yann LeCun and Yoshua Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, 1995.

[8] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[9] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, and Gerald Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4277–4280.

[10] Tara N Sainath, Abdel-rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran, "Deep convolutional neural networks for lvcsr," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8614–8618.

[11] Rasmus Berg Palm, "Prediction as a candidate for learning deep hierarchical models of data," *Technical University of Denmark, Palm*, 2012.

[12] R.W. Schafer, "What is a Savitzky-Golay filter? [lecture notes]," *Signal Processing Magazine, IEEE*, vol. 28, no. 4, pp. 111–117, 2011.

[13] Francesco Beritelli, Salvatore Casale, and A Cavallaero, "A robust voice activity detector for wireless communications using soft computing," *Selected Areas in Communications, IEEE Journal on*, vol. 16, no. 9, pp. 1818–1829, 1998.