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

Ultrasonic Sensing for Robust Speech Recognition

Sundararajan Srinivasan, Bhiksha Raj, Tony Ezzat

TR2010-015 April 2010

Abstract

In this paper, we present our work using ultrasonic sensing of speech for digit recognition. First, a set of spectral ultrasonic features are developed and tuned in order to achieve optimal performance for the digit recognition task. Using these features, we demonstrate an overall accuracy of 33.00% on a digit recognition task using HMMs with recordings from 6 speakers. The results indicate that ultrasonic sensing of speech is viable, but that further work is needed to achieve word accuracies that match those of audio. Finally, experimental results are presented which demonstrate that fusing information from ultrasound and audio sources show marginal improvements over audio-only performances.

ULTRASONIC SENSING FOR ROBUST SPEECH RECOGNITION

Sundararajan Srinivasan
Department of Electrical
and Computer Engineering
Mississippi State University
ss754@ece.msstate.edu

Bhiksha Raj
Associate Professor
Language Technologies Institute
Carnegie Mellon University, PA
bhiksha@cs.cmu.edu

Tony Ezzat
Mitsubishi Electric
Research Laboratories
Cambridge, MA.
ezzat@merl.com

ABSTRACT

In this paper, we present our work using ultrasonic sensing of speech for digit recognition. First, a set of spectral ultrasonic features are developed and tuned in order to achieve optimal performance for the digit recognition task. Using these features, we demonstrate an overall accuracy of 33.00% on a digit recognition task using HMMs with recordings from 6 speakers. The results indicate that ultrasonic sensing of speech is viable, but that further work is needed to achieve word accuracies that match those of audio. Finally, experimental results are presented which demonstrate that fusing information from ultrasound and audio sources show marginal improvements over audio-only performances.

Index Terms— ultrasound, digit recognition, fusion

1. INTRODUCTION

Ultrasound Doppler sensing of speech offers a new a exciting paradigm for sensing speech. A pure tone in the ultrasound range is emitted by a transmitter facing the speaker, and the reflected wave captured by a receiver. The reflected wave undergoes frequency "Doppler" shifts and amplitude envelope modulations proportional to lip and facial mouth movements. By analyzing the received ultrasound signal we hope to decode the underlying sounds associated with speech.

Ultrasound sensing of speech is important in cases where the speaker needs a "silent speech interface", i.e., when speech needs to be uttered in an inaudible way. These cases arise when the speaker is in a public space and does not wish to be heard speaking aloud, or when public etiquette dictates that the speaker not annoy those around him/her. Additionally, ultrasound sensing of speech is important as an additional modality to audio itself, by providing additional features to a recognizer that encode information about the movement of the articulators. Since the ultrasound signal is not corrupted by additive noise in the audible range it is expected that the ultrasonic signal will be robust in low SNR's.

All previous work involving ultrasonic sensing of speech have either not addressed speech recognition, used unrealistic settings, or otherwise lack reproducibility due to the use of non-standard recognition architectures. in [2][3], Ultrasonic Doppler sensing of speech is employed for speaker recognition and voice activity detection. a more exhaustive analysis of ultrasonic sensing for speech recognition was done in [4]. However, this work used a segmental landmark-based recognizer instead of a more standard frame-based HMM framework. Additionally, in [4] an artificial constraint on the number of digits in the utterance was imposed using a fixed-length loop grammar.

In this work, we sought to perform ultrasonic sensing of digits using the standard AURORA HTK architecture [5][6], in which the ultrasonic features are used to train standard left-to-right digit HMMs. This framework allows for more direct comparison with audio, and enables reproducibility of our results on a common platform used by many other researchers. Additionally, we imposed no constraint on the number or digits in each utterance (as is done in AURORA).

Another motivation for our work is to explore the performance of ultrasonic sensing of speech when the sensor is located at 16inches from the speaker. In previous work [4], the sensor was at 6-8 inches from the face. In many applications, such as when the sensor is affixed to a car dashboard, computer screen, or kiosk, a distance of 16 inches is a more realistic distance setting than 6-8 inches.

The outline of our paper is as follows: in section 2, a brief description of the ultrasound board we used is provided. Section 3 provides details of our ultrasonic features and tuning experiments. In section 4, the data set we use for recognition and evaluation is provided. In section 5, results from ultrasound-only experiments are shown and discussed. in section 6, results from our fusion experiments with audio are shown and discussed. Finally, we offer conclusions and suggestions for future work in section 6.

2. ULTRASOUND BOARD DESCRIPTION

The basic building blocks of the ultrasound board used in this work are depicted in Fig 1. the board is identical to the one used in [4][7]. The main advantage of this hardware over those used in prior work is that it enables synchronized two-channel output of audio and ultrasound

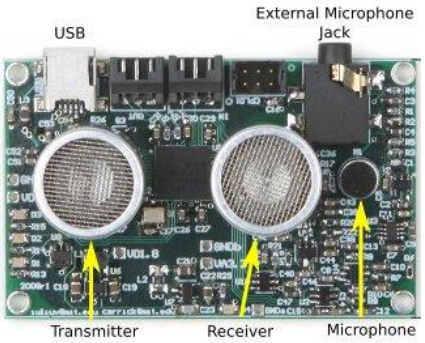

Fig. 1. Picture of the Ultrasound Board

signals. The microphone is responsible for the audio capture. The ultrasound transmitter emits a 40 kHz carrier tone and the receiver captures the reflected and modulated carrier. The ADC in the board delivers ultrasound samples at 24 kHz rate, hence the carrier gets aliased to 8 kHz. Audio signal is sampled at 16 kHz in synchrony with ultrasound. An example spectrogram of recorded ultrasound signal and a time-slice of the spectrogram are shown in Fig. 2. It is clear from this figure that information is concentrated around the carrier. After modulation, the information in speech would be encoded in the frequency shifts and envelope amplitude variation of the carrier.

3. ULTRASOUND FEATURE EXTRACTION AND PARAMETER TUNING

An appropriate set of ultrasound features were developed before recognition experiments were performed. The features are cepstral in nature, and their parameters were chosen after a round of tuning experiments designed to optimize recognition accuracy on a development set consisting of two speakers who spoke at a distance of 16inches from the board.

Before feature extraction is performed, the ultrasound signal is first pre-processed: the signal is zero-meaned, then

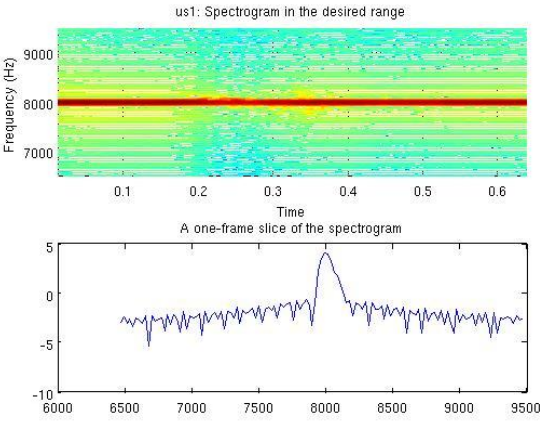

Fig. 2. Spectral Information in the Ultrasound Signal

passed through a MA filter that averages three consecutive samples to suppress the 8khz carrier. Finally, the MA'd signal is passed through a difference operator. The frequency response of this processing is shown in Fig 3. A spectrogram of the pre-processed signal is shown in Fig 4.

Frames of 50 msec are extracted from the pre-processed signal via a Hamming window. Then a 2048-pt FFT is performed per frame. The FFT bins corresponding to the 7-Khz frequency range are extracted, the logarithm of their magnitudes taken, and finally, a DCT is applied to compact energy in the first few coefficients and de-correlate the signal. Tuning experiments revealed that the first 38 DCT coefficients along with energy offered a reasonable set of static features. Finally, velocity and acceleration coefficients are appended yielding a final vector length of 117 features.

The results of our tuning experiments to fix the frequency range of interest are shown in Table I. From this we note that 7 kHz to 9.5 kHz is about optimal.

We also need to tune the ultrasound gain and ultrasound power. The former refers to the sensitivity of the ultrasound receiver and the latter to the strength of the tone the transmitter emits. For tuning gain, we fixed power at 3 and noted the performance variation with varying gain. The

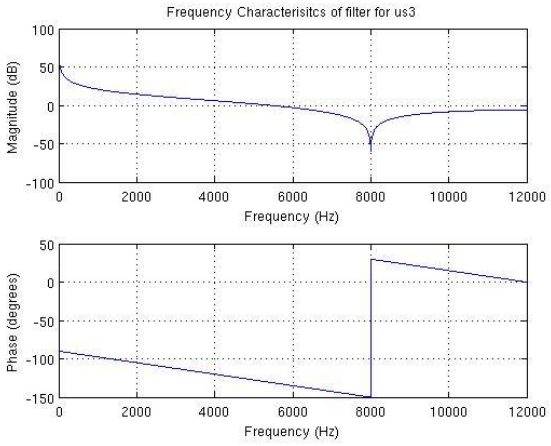

Fig. 3. Frequency Response of the Ultrasound Preprocessing Filter

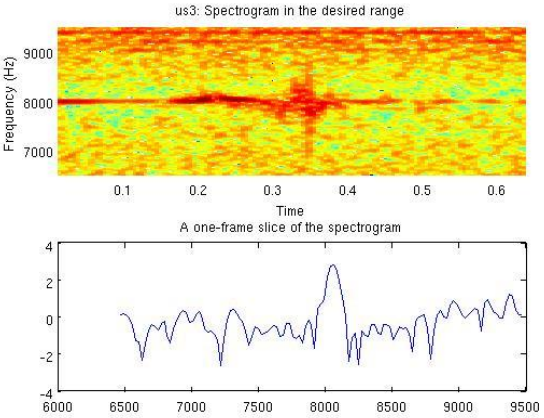

Fig. 4. Spectrogram and one time-slice of it of the preprocessed ultrasound in the frequency range of interest for utterance "*one*".

Table I
Tuning Frequency Range
for Ultrasound Feature Extraction

Freq range	Fast speech	Slow speech
4.0-12.0	39.33	53.33
5.5-9.0	39.33	68.00
6.0-9.5	43.33	60.00
6.5-9.5	34.67	70.67
7.0-10.0	42.00	72.00
7.0-9.5	44.00	80.00
7.0-9.0	42.00	77.86
7.5-9.0	36.67	76.67
7.5-8.5	32.67	60.00

results are in Table II. From this, we fixed gain value at 5. For tuning power, we fixed gain at 5 and noted the performance variation with varying power. The results are in Table III. From this we fixed the power value at 3.

4. DESCRIPTION OF DATASET

We recorded 40 utterances from 6 male speakers each. The ultrasound board is capable of recording both audio and ultrasound signal in synchrony. Each utterance had all 10 digits in the set {ZERO, ONE, …, NINE} in random order and without repetition. Since speed appeared to affect the ultrasound performance, for each speaker, 20 utterances were uttered slowly with pauses in-between of more-or-less equal duration, while the other 20 were uttered quite fast in 3-3-4 format. The final ultrasound board settings we used after tuning are: distance of the board from the speaker=16 in., ultrasound gain = 5, ultrasound power = 3.

Also, to simulate different audio noise conditions, we artificially add noise recorded in a Mercedes car to the audio signal. To corrupt a given audio signal, a random segment of the same length as the audio signal is extracted from the car noise. Then, depending on the SNR level, the noise segment is scaled and added to the audio signal. This was achieved using the FaNT software that is popular for this purpose [6].

5. EXPERIMENTS WITH ULTRASOUND DATA

We first investigate the performance of ultrasound features for speaker-independent digit recognition. Our setup is similar to the AURORA experimental framework [5] using HTK [6]. Since we only have a limited amount of data – 6 speakers – we used leave-one-out strategy for evaluation, i.e., we evaluate all 40 utterances from each speaker using models trained using all the remaining 5 speakers, and then state the average over all the 6 speakers as the final digit recognition accuracy. All digit models are 16-state HMMs

Table II
Tuning Ultrasound Gain
(with Ultrasound Power fixed at 3)

Gain	Fast speech	Slow speech
4	24.67	37.33
5	43.33	54.00
6	30.00	30.00

Table III
Tuning Ultrasound Power
(with Ultrasound Gain fixed at 5)

Power	Fast speech	Slow speech
2	30.67	43.33
3	35.33	57.14
4	39.33	20.67
5	38.67	19.33

with one Gaussian per state. Only one mixture was used due to limited amount of data.

Our results are shown in Table 4. Using ultrasound-only features we observed a mean speaker-independent digit recognition accuracy of 33%. While this is low compared to the audio-only performance of 94.79%, even this high a performance is still surprising and very encouraging. Our results compare favorably to the results of [4] which indicated WER of 70.5% using ultrasound alone, even though [4] was constraining the recognition to exactly 10 digit strings.

6. FUSION OF ULTRASOUND AND AUDIO INFORMATION

We also investigated the effect of fusion of information from ultrasound and audio signals on speech recognition performance over a variety of noise conditions. To achieve this, we use a asynchronous decision fusion technique [1] to fuse the scores from the two models: independent models are trained for audio and ultrasound; n-best lists are generated from audio and rescored using both audio and ultrasound; then the hypothesis with the highest weighted combination is chosen. We chose this technique for fusion because of the observed asynchronicity between audio and ultrasound signal (events in ultrasound signal precede those in audio by hundreds of milliseconds).

An example of variation in digit accuracy with varying alpha is shown in Table V using models trained using clean data and evaluated with noisy data with SNR=5dB. From

Table IV
Ultrasound- and audio- only performances (accuracy %)

Mode	Fast speech	Slow speech	Both
Audio	93.67	78.92	94.79
Ultrasound	37.75	18.17	33.00

Table V
Variation of Digit Accuracy with Combination Weight α. Clean Train – Noisy (SNR=5 dB) test and using audio 10-best list.

α	Accuracy %
0.0	60.67
0.1	60.58
0.2	60.96
0.3	61.42
0.4	61.46
0.5	61.33
0.6	61.08
0.7	60.67
0.8	60.75
0.9	60.96
1.0	60.75

this, we see that there is a 2.71% absolute improvement in accuracy at α=0.5 as compared to using audio alone (α=0.0).

Table VI shows the performance after fusion for evaluation data at various SNRs. The clean-train condition indicates that the corresponding audio models are trained using only clean audio. This should provide an estimate of performance in unseen noise conditions. Multiconditioned-train condition indicates that the models are trained with training data from all those SNRs. From table 5 we see that average performance gain by adding ultrasound in clean train condition is 0.46% absolute, and for multiconditioned training is 0.09% absolute. In both these conditions, we see that addition of ultrasound information to audio improves performance only marginally. Our results differ markedly from those in [4] where much stronger performance gains are obtained by using ultrasound in low SNR conditions (6-15% absolute improvement for SNR levels of 10dB and 0db). It is unclear whether the performance gains reported there are due to better feature extraction, lower distance of the speaker to the board, the artificial constraint on the number of digits in the utterance imposed using a fixed-length loop grammar, or a better fusion technique.

7. CONCLUSIONS AND FUTURE WORK

In this work, we have explored the use of ultrasound Doppler sensing of speech for speech recognition. Some details of the parameter tuning experiments and the feature extraction technique were mentioned. Significant variation in performance with utterance speed was noted. This deserves further exploration in the future to normalize performance across different utterance speeds and hence to increase the overall performance. A WER of 33% was observed on digit recognition using ultrasound alone. A marginal improvement in digit accuracy was shown by fusing information in ultrasound signal with that of audio.

Table VI
Performances of fusion experiments (accuracy %).

SNR (Noisy Test)	With Clean Train		With Multiconditioned Train	
	AU-only	Fusion	AU-only	Fusion
Clean	94.79	95.25	94.58	94.67
10dB	77.21	78.04	90.21	91.00
5dB	60.67	61.46	86.21	87.12
0dB	36.75	36.75	74.08	76.04

Further work on more efficient fusion of the two information sources is also necessary before ultrasonic sensing can be considered in practical multimodal speech recognition systems. Finally, other aspects that require attention are the variation of ultrasound performance with changes in pose (the angle between the ultrasound board and the face of the speaker), distance of the speaker from the board and Doppler noise due to wind effects.

8. ACKNOWLEDGEMENTS

The authors would like to thank Dr. James Glass, Carrick Detweiler and Iuliu Vasilescu for providing us the ultrasound capture board used in this research.

9. REFERENCES

[1] Audio-Visual Speech Recognition. C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari and J. Zhou. *Technical Report*, Workshop 2000, CLSP, Johns Hopkins University, July-August 2000.

[2] K. Kalgaonkar and B. Raj, "Ultrasonic doppler sensor for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2008, pp. 4865–4868.

[3] K. Kalgaonkar and B. Raj, "Ultrasonic doppler sensor for voice activity detection," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 754–757, October 2007.

[4] B. Zhu, "Multimodal speech recognition with ultrasonic sensors," *Master's thesis*, Massachusetts Institute of Technology, Cambridge, Massachusetts, September 2008.

[5] D. Pearce and H. Hirsch, The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions, *ICSLP 2000*, pp. 29-32, Oct. 2000.

[6] Hidden Markov model Toolkit ver. 3.4.1, available from http://htk.eng.cam.ac.uk.

[7] C. Detweiler and I. Vasilescu, "Ultrasonic speech capture board: Hardware platform and software interface," Independent study final paper, MIT, May 2008.

[8] Filtering and Noise-adding Tool (FaNT), available from http://dnt.kr.hsnr.de/download.html