



Multimodal Speech Recognition with Ultrasonic Sensors

Bo Zhu, Timothy J. Hazen, and James R. Glass

MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA USA

{boz, hazen, glass}@csail.mit.edu

Abstract

In this research we explore multimodal speech recognition by augmenting acoustic information with that obtained by an ultrasonic emitter and receiver. After designing a hardware component to generate a stereo audio/ultrasound signal, we extract sub-band ultrasonic features that supplement conventional MFCC-based audio measurements. A simple interpolation method is used to combine audio and ultrasound model likelihoods. Experiments performed on a noisy continuous digit recognition task indicate that the addition of ultrasonic information reduces word error rates by 24-29% over a wide range of acoustic SNR (20-0 dB).

Index Terms: multimodal, ultrasonic speech recognition

1. Introduction

Conventional automatic speech recognition (ASR) engines use the recorded audio signal as the sole source of input information. It has long been known however, that humans can get additional performance gains from parallel information available from the talker (e.g., visual), especially when the audio channel is corrupted by noise [1]. This observation has motivated a significant number of investigations that have explored combinations of audio with information garnered from other modalities. This research has studied both alternative modes and their feature representations, as well as architectures for combining parallel sources of loosely-coupled information [2, 11].

Some multimodal research has examined methods that can be used in “tethered” scenarios, whereby the subject can have devices physically touching their head, face and/or throat [4]. Example applications include pilots in a noisy cockpit augmenting audio with throat and nose recordings [10], as well as headsets that use a bone-conducting microphone to reduce the effect of environmental noise [14].

Other multimodal research has examined “untethered” scenarios where the subject is not physically connected to the recording devices. There are a wide range of potential application areas for such technologies including PDAs, kiosks, vehicles, etc. Due to the potential increased noise and reverberation resulting from distant audio recordings, these scenarios can potentially benefit from multimodal processing methods. One of the more commonly examined areas of multimodal research is audio-visual speech recognition (AVSR) [2, 5, 9, 11]. AVSR research has demonstrated significant ASR performance gains, especially in noisy acoustic environments. However, there are other modalities that have been considered that are potentially less expensive than visual-based processing, or might be more acceptable to users who do not want to be visually recorded. For example, there has been research using micro-pulse radar based techniques that can measure vocal-fold vibration in a non-invasive manner [8]. Other researchers have explored ultrasonic sensors to complement audio-based recordings for use in speech

detection [7] and recognition [6].

In this work, we describe our preliminary work in attempting ultrasonic speech recognition. This work differs from earlier ultrasonic ASR efforts [6] in that we are using a statistical speech recognizer and a continuous speech recognition task. It also differs from more recent ultrasonic research that has focused primarily on robust voice activity detection [7].

In the following sections we first provide background on the basics of ultrasonic processing, and then describe the ultrasonic hardware that was developed for this project. We then describe the acoustic features we chose to extract from the ultrasonic signal. Finally, we describe our ultrasonic speech recognition experiments for a continuous digit task, where we found word error rate reductions from 24% to 29% over a range of noisy audio conditions (20-0dB SNR).

2. Background

Ultrasonic-based processing is performed in a similar manner to radar, whereby an ultra high-frequency acoustic tone (e.g., 40 kHz) is directed at a moving object, causing reflections which are recorded by a receiver which is usually co-located with the emitter. The frequency of the reflected tone will be governed by the Doppler effect [13], and can be expressed as $f = f_0(1 + \frac{v}{c})$, where f_0 is the frequency of the emitted tone, f is the frequency of the reflected tone, v is the velocity of the reflecting surface towards the emitter, and c is the velocity of sound. Thus, if the ultrasonic tone reflects off of a surface moving towards the emitter, the received signal will have a higher frequency. Likewise, a lower frequency tone will be recorded when the reflective surface is moving away from the emitter.

In the case where the ultrasonic beam is reflecting off of a complex moving object (e.g., a talkers face), there will be many different reflections recorded by the receiver. In general, the reflected signal will be a sum of sinusoids of varying strengths and frequencies. The time-varying patterns of these reflected signals provide information about the nature of the motion. In the case of a talkers face for example, reflections will be caused by articulator motion during speech production, as well as other motion that reflects the ultrasonic beam (e.g., head, body). Raj and his colleagues have shown that the ultrasonic time-frequency patterns of speech are distinct from many other kinds of motion, and can be used for robust voice activity detection [7]. However, the time-frequency patterns can potentially be useful for discriminating among speech sounds themselves.

Figure 1 shows a regular spectrogram along with an ultrasonic spectrogram for the utterance “ma na”. The ultrasonic spectrogram shows perturbations from the carrier signal at 4.4 kHz that capture motion of the articulators at the release of both consonants. The effect is measurable even though the major motion is parallel to the ultrasonic beam. Moreover, it is interesting to observe that the ultrasonic signature associated with

10.21437/Interspeech.2007-284

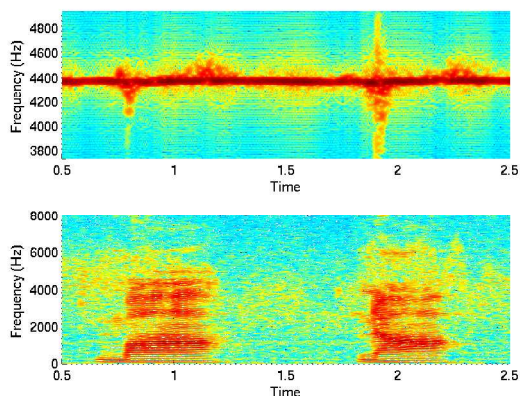


Figure 1: Ultrasonic (top) and audio (bottom) spectrograms of the utterance “ma na”.

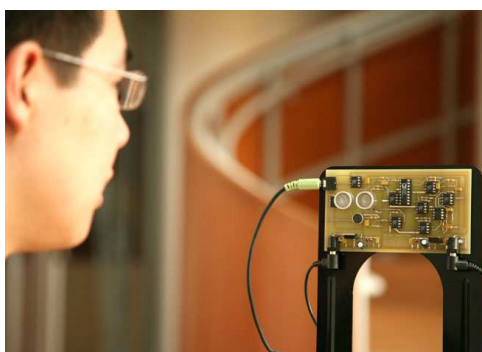


Figure 2: User speaking into the hardware capture device. On-board electronics prepare the microphone and ultrasonic signals for sound card data acquisition.

the release of the consonants tends to be associated with the place of articulation. Thus, an ultrasonic signal can potentially add valuable information to the acoustic signal, especially in noisy ASR environments.

3. Hardware

Before we could begin ASR experiments, we needed to create a hardware component that would generate and receive an ultrasonic signal. In order to simplify the need for any additional equipment, we decided to also include an audio microphone, so that we could create a stereo acoustic signal that could be input to conventional computers with an on-board A/D device. Since a 40 kHz carrier tone is higher than the largest sampling rate of most conventional A/D converters, we decided to frequency modulate the spectrum of the ultrasonic signal so that the stereo signals could be sampled at a sampling frequency of 16 kHz/s. The resulting hardware that we developed is shown in Figure 2. The primary sensors it contains are an ultrasonic emitter and receiver, and an electret microphone for the regular audio signal.

The ultrasonic transmitter is a Kobitone 400ST160 tuned to a resonant frequency of 40 kHz. The transmitter is driven by a 40 kHz squarewave generator, which is implemented by a PIC10F206 microcontroller. The output of the transmitter is a pure sinusoid even though it is driven by a squarewave, because the transmitter is inherently a narrowband device that will bandpass filter the other harmonics, leaving the first 40 kHz sinusoidal harmonic.

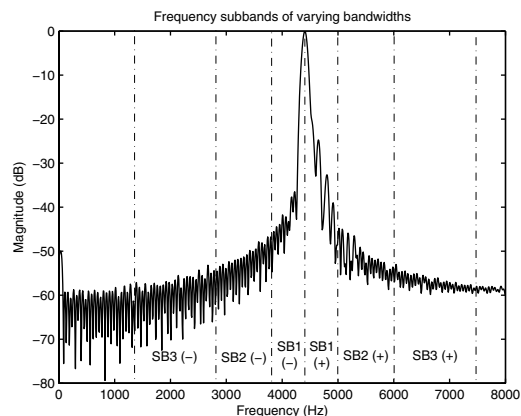


Figure 3: Example of six frequency sub-bands on an ultrasonic spectral slice. Average energy is computed for each sub-band.

The ultrasonic receiver is a Kobitone 400SR160 also centered around 40 kHz, with a -6dB bandwidth of 2.5 kHz. This bandwidth allows minor frequency shifts to be detected, and these frequency shifts are the basis of our subsequent analysis. In order to shift the ultrasonic spectrum down to a lower frequency range, the received signal is frequency modulated with a 35.6 kHz sinusoid to downshift it to be centered at 4.4 kHz, well within the capture bandwidth of standard sound cards. The modulation process is implemented by a 35.6 kHz squarewave generator (also a PIC10F206) and a fourth-order butterworth lowpassfilter with a cutoff frequency at 48 kHz. This cutoff frequency eliminates the odd harmonics above the first, resulting in a 35.6 kHz sinusoid. An Analog Devices MLT04 analog multiplier is then used to multiply the received signal and the sinusoid to perform the modulation.

4. Feature Extraction

As described earlier, the recorded ultrasonic signal will consist of a number of different frequency components, with each component corresponding to a reflection from a moving (articulator) surface. The amount of energy associated with a particular frequency (relative to the carrier frequency) can be associated with articulator(s) moving with a certain velocity at a particular time. We therefore tried to extract simple acoustic measurements that would capture the distribution in spectral energy as a function of time.

Our first measurements, as illustrated in Figure 3, partitioned the ultrasonic spectrum into fourteen non-linearly-spaced sub-bands centered around the carrier frequency of 4.4kHz. The bandwidths slowly increased from 40 Hz to 310 Hz from the first to the seventh band, respectively. These sub-bands were a crude attempt to measure the amount of energy (relative to the carrier tone) in different portions of the spectrum. The non-linear spacing was an attempt to be more sensitive to portions of the spectrum near the carrier frequency.

The second set of measurements, as illustrated in Figure 4, attempted to quantify frequency deviation from the center frequency in different parts of the spectrum, as could be observed in the ultrasonic spectrogram. This was accomplished by measuring the center of mass (COM) in frequency regions bounded by particular energy thresholds relative to the energy of the carrier frequency. Several energy thresholds were used to compute a variety of COMs over ranges: 0-20 dB, 20-40 dB, 40-60 dB

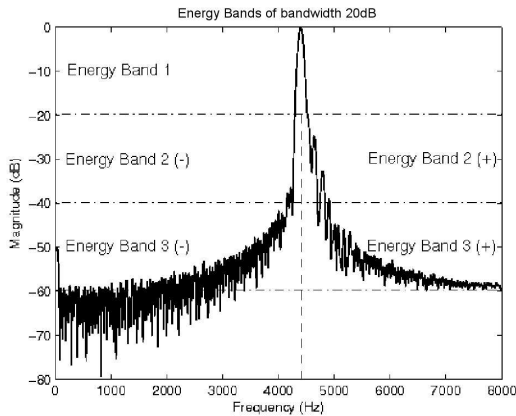


Figure 4: Example of five energy sub-bands on an ultrasonic spectral slice. Center-of-mass calculations are performed over frequency ranges defined by relative energy thresholds.

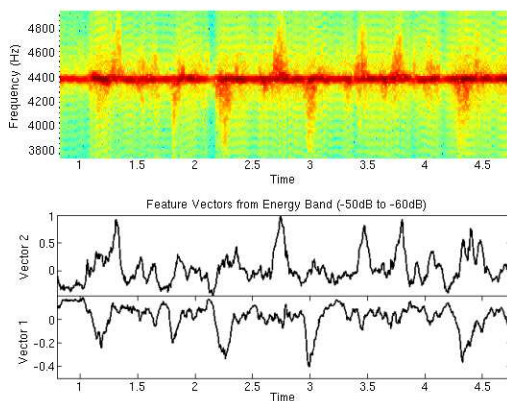


Figure 5: Ultrasonic spectrogram of utterance “4 6 7 0 5 0 5 5 6 0” and two feature vectors obtained from the energy sub-band between -50dB and -60dB.

etc. Thirteen total frequency centroids were computed for each frame. Figure 5 demonstrates the feature vectors from the energy band -50dB to -60dB closely following the outside envelope of the spectrogram.

5. Experimental Setup

5.1. Data Collection

Data collection was performed at MIT in quiet office environments. The corpus consisted of twenty talkers: nineteen male and one female. The talkers were situated in front of the ultrasonic transducers, with a distance of about six inches between the talker’s face and the transducers. The talkers were prompted with fifty sequences of ten randomized digits each. The digits were 0 through 9, and the users were told to say “zero” instead of “oh” for consistency. The entire data set consisted of one thousand ten-digit utterances; each digit was spoken approximately one thousand times. For our experiments, we divided our collected data into a training set containing 750 utterances from 15 speakers, and a test set containing 250 utterances from a disjoint set of 5 speakers.

5.2. Speech Recognition Configuration

Our speech recognition experiments were conducted using our landmark-based speech recognizer that has been previously used for AVSR experiments [3, 5]. The recognizer was configured to recognize random digit strings containing exactly 10 digits. The digit strings were modeled by 110 context-dependent diphone-based acoustic and ultrasonic models.

To generate the landmark-based acoustic features, the speech signal is first processed into frame-based Mel-frequency scale cepstral coefficients (MFCCs) at a rate of 200 frames per second. Each frame consists of a vector of 14 MFCCs. From the MFCC frames, significant landmarks in the acoustic signal are first detected using a measure of acoustic change. Feature vectors are extracted at landmarks based on averages of MFCC vectors in the region surrounding each landmark. Specifically, a set of 8 telescoping regions are defined which together span 150ms around the landmark. Within each of these regions the frame-based MFCC feature vectors are averaged to form a single 14-dimension feature vector for the entire region. In total this yields a single 112-dimension (8 regions \times 14 dimensions) feature vector for each landmark. The landmark feature vectors are then projected down to 50-dimensions using principle components analysis. From the 50-dimension feature vectors, word-dependent diphone-based phonetic models are created to represent the acoustic landmarks within the digit words. Mixture Gaussian density functions were used to model the 110 different diphone models.

The models capturing the ultrasonic measurement information were generated in a similar fashion as the acoustic models. For every frame the ultrasonic signal is represented by the collection of 27 ultrasonic measurements (13 energy-band frequency centroids and 14 frequency sub-band energy averages). Within each of six telescoping regions surrounding an acoustic landmark, the ultrasonic frame vectors are averaged to form a single 27-dimension feature vector for the entire region. The full set of six regions spans 140ms around the landmark. In total, this yields a 162-dimension (6 regions \times 27 dimensions) ultrasonic feature vector for each landmark. The ultrasonic landmark feature vectors are then also projected down to 35-dimensions using principle components analysis. As with the acoustic information, the ultrasonic information is modeled with a mixture Gaussian density function for each of 110 different diphone models.

In addition to the acoustic and ultrasonic models, a context-independent phonetic duration model was also created. The three models were trained on the data in the 15 speaker training set. In the baseline recognizer configuration, the acoustic, ultrasonic and duration models were combined with equal weights of 1. In situations where there may be considerable background acoustic noise, the system can reduce the weight of the acoustic model relative to the ultrasonic model as the acoustic signal-to-noise ratio (SNR) is reduced.

To simulate noisy acoustic conditions, babble noise from the NOISEX database was synthetically added to the data in the test set at SNR levels of 20db, 10db and 0db [12]. This provided us with four noise conditions (including the clean condition) for our experiments. At each noise condition we examined the recognition performance as the weight of the acoustic model was varied from 0.0 to 1.0.

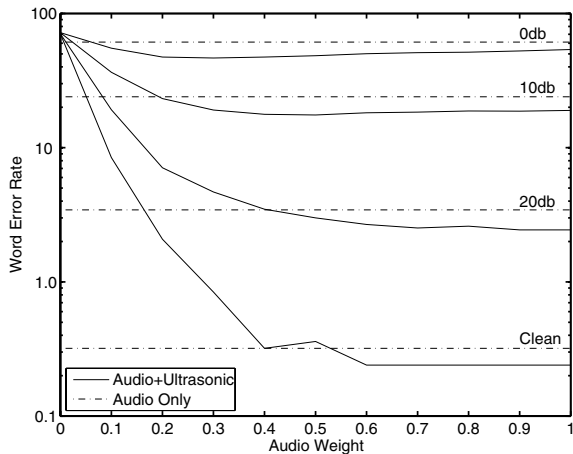


Figure 6: Speech recognition results for the multimodal audio+ultrasonic recognizer for four noise levels as the audio weight is varied from 0.0 to 1.0. The result of the audio only recognizer is also shown for each of the four noise levels.

Noise Level	Optimal Audio Weight	Word Error Rate (%)		
		Audio Only	Ultrasonic Only	Audio + Ultrasonic
Clean	1.0	0.32	70.5	0.24
20db	1.0	3.44	70.5	2.44
10db	0.5	24.0	71.8	17.5
0db	0.3	61.2	72.0	46.6

Table 1: Digit recognition results for the audio-only, ultrasonic-only, and multimodal (audio+ultrasonic) systems when the optimal audio weight is used.

6. Results

Figure 6 shows a graph containing our full set of results. A solid curve for each noise condition shows the multimodal (audio+ultrasonic) recognition results as the audio weight is varied from 0.0 to 1.0, and a dashed-line is shown for the unimodal audio-only result at each noise level. The graph shows that the ultrasonic information improves the speech recognition result over the audio-only case for a wide range of audio weights for each condition. Table 1 summarizes the best results from the figure, i.e., the minimum error rates that were obtained from the multimodal system at the optimal audio weight setting. Over the four different noise conditions, error rate reductions from audio-only to the audio+ultrasonic system varied between 24% and 29% at the optimal audio rate setting.

7. Discussion and Future Work

The combination of audio and ultrasonic signals has been shown to be effective in building a more noise-robust speech recognizer. This type of system is advantageous in any situation where the user is at a reasonably close distance from the sensors. Kiosks in building lobbies, navigation systems in cars or airplane cockpits, and automated systems on loud factory floors can all benefit from an audio+ultrasonic ASR. It is worth noting that the improvements obtained from adding the ultrasonic information to our recognizer are comparable to improvements we have observed in experiments that incorporate visual lip-reading into our recognizer [5]. This is particularly noteworthy because the processing of the ultrasonic signal performed in our experiments is considerably less complex than the

visual processing required to perform visual lip-reading. Moreover, ultrasonic recordings are not as sensitive to some users as visual-based recordings.

In future work we plan to explore alternative feature extraction methods and more rigorously test the ultrasonic device by deploying it in a publicly available kiosk on a medium-vocabulary task. This will allow us to measure sensitivity to talker location and speaking style.

8. Acknowledgements

The authors would like to thank Bhiksha Raj and Kaustubh Kalgaonkar for their help in creating the ultrasonic hardware component used in this research.

9. References

- [1] L. Bernstein, and C. Benoit, "For speech perception by humans or machines, three senses are better than one," *Proc. ICSLP*, Philadelphia, 1996.
- [2] S. Chu and T. Huang, "Audio-visual speech modeling using coupled hidden Markov models," *Proc. ICASSP*, Orlando, 2002.
- [3] J. Glass, "A probabilistic framework for segment-based speech recognition," *Computer, Speech, and Language*, 17(2-3), 137–152, 2003.
- [4] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, "Combining standard and throat microphones for robust speech recognition," *IEEE Signal Processing Letters*, 10(3), 72–74, 2003.
- [5] T. J. Hazen. "Visual model structures and synchrony constraints for audio-visual speech recognition," *IEEE Trans. ASLP*, 14(3), 1082–1089, 2006.
- [6] D. L. Jennings and D. W. Ruck, "Enhancing automatic speech recognition with an ultrasonic lipmotion detector," *Proc. ICASSP*, Detroit, 1995.
- [7] K. Kalgaonkar and B. Raj, "An acoustic doppler-based front-end for hands free spoken user interfaces," *IEEE/ACL Workshop on SLT*, Aruba, 2006.
- [8] C. Kwan, X. Li, D. Lao, Y. Deng, Z. Ren, B. Raj, R. Singh, and R. Stern, "Voice driven applications in non-stationary and chaotic environment," *Trans. IJSP*, 3(4), 2006.
- [9] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, 91(9), 1306–1326, 2003.
- [10] S. Roucos, V. Viswanathan, C. Henry, and R. Schwartz, "Word recognition using multisensor speech input in high ambient noise," *Proc. ICASSP*, Tokyo, 1986.
- [11] S. Tamura, K. Iwano and S. Furui, "A stream-weight optimization method for audio-visual speech recognition using multi-stream HMMs," *Proc. ICASSP*, Montreal, 2004.
- [12] A. Varga and H. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, 12(3), 247–251, 1993.
- [13] http://en.wikipedia.org/wiki/Doppler_effect
- [14] Z. Zhang, et al., "Multi-sensory microphones for robust speech detection, enhancement and recognition," *Proc. ICASSP*, Montreal, 2004.