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

Ultrasonic Doppler Sensor For Speaker Recognition

Kaustubh Kalgaonkar, Bhiksha Raj

TR2008-014 August 2008

Abstract

In this paper we present a novel use of an acoustic Doppler sonar for multi-modal speaker identification. An ultrasonic emitter directs a 40kHz tone toward the speaker. Reflections from the speakerś face are recorded as the speaker talks. The frequency of the tone is modified by the velocity of the facial structures it is reflected by. The received ultrasonic signal thus contains an entire spectrum of frequencies representing the set of all velocities of facial components. The pattern of frequencies in the reflected signal is observed to be typical of the speaker. The captured ultrasonic signal is synchronously analyzed with the corresponding voice signal to extract specific characteristics that can be used to identify the speaker. Experiments show that the information this can result in significant improvements in speaker identification accuracy both under clean conditions and in noise.

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Copyright © Mitsubishi Electric Research Laboratories, Inc., 2008
201 Broadway, Cambridge, Massachusetts 02139

ULTRASONIC DOPPLER SENSOR FOR SPEAKER RECOGNITION

*Kaustubh Kalgaonkar**

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332
kaustubh@ece.gatech.edu

Bhiksha Raj

Mitsubishi Electric Research Laboratories
Cambridge, MA 02139
bhiksha@merl.com

ABSTRACT

In this paper we present a novel use of an acoustic Doppler sonar for multi-modal speaker identification. An ultrasonic emitter directs a 40kHz tone toward the speaker. Reflections from the speaker's face are recorded as the speaker talks. The frequency of the tone is modified by the velocity of the facial structures it is reflected by. The received ultrasonic signal thus contains an entire spectrum of frequencies representing the set of all velocities of facial components. The pattern of frequencies in the reflected signal is observed to be typical of the speaker. The captured ultrasonic signal is synchronously analyzed with the corresponding voice signal to extract specific characteristics that can be used to identify the speaker. Experiments show that the information this can result in significant improvements in speaker identification accuracy both under clean conditions and in noise.

Index Terms— Speaker Recognition, Speaker Verification, Ultrasonic Doppler Sensor.

1. INTRODUCTION

The problem of speaker identification has traditionally been treated as one of *audio* classification, e.g. [1, 2]. The speech from the speaker is parameterized into sequences of feature vectors. The sequences of feature vectors are classified as belonging to a particular speaker using some classification mechanism. Research has primarily focused either on deriving newer and more descriptive features from the audio [3, 4], on the classification mechanisms and models employed [5, 6]

It has lately been recognized that speaker identification performance can be greatly enhanced by augmenting measurements from the speech signal with input from other sensors, in particular a camera. A variety of techniques have been proposed to integrate information extracted from the video with that obtained from the audio. The most obvious is to combine evidence from a face-recognition classifier that operates on the video to evidence from the speaker ID system that works on the audio [7]. Other techniques have explicitly to derive speaking-related features, such as characterizations of lip configurations, facial texture around the lips [8] etc.

Other secondary sensors such as Pmics and GEMS sensors have been been proposed in the literature to provide measurements that augment audio signals; however they have largely been used for speech *recognition*, since they primarily produce readings that represent relatively noise-free readings of the some aspects of the speech signal (such as a filtered version of the speech, or the excitation to the vocal tract) and do not provide any additional information about the speaker that is not contained in the speech signal itself. Additionally, many of them must be mounted on the person of the speaker and are not appropriate for use in most speaker identification/verification scenarios.

In this paper we propose the use of an entirely different type of secondary sensor for speaker ID – an acoustic Doppler sonar (ADS). The ADS is an inexpensive far-field sensor that can obtain measurements of movements of a talker's face. The ADS has previously been used successfully for voice activity detection [9]. It has also been shown that Doppler readings can be effectively used as secondary measurements to improve automatic speech recognition in noise [10]. Here we show that they can be effectively used for speaker identification as well.

The ADS consists of a high-frequency ultrasound emitter and an acoustic transducer that is tuned to the transmitted frequency. An ultrasound tone output by the emitter is reflected from the speaker's face and undergoes a Doppler frequency shift that is proportional to normal velocity of the portion of the face that it is reflected by. The reflected "Doppler" signal thus contains an spectrum of frequencies that represent the motion of the speakers cheeks, lips, tongue, etc. The pattern of movements of facial muscles while speaking is typical of the talker. By characterizing the velocities of these movements, the Doppler signal thus represents a signature that is quite specific to the person. Importantly, the information present in the Doppler signal is not directly also represented in the audio itself.

In Section 2 we describe the basic hardware setup of the ADS. Our setup, built with off-the shelf components, costs only a few dollars (US); if replicated on a large scale it can be made far cheaper. In Section 3 we briefly discuss the Doppler principle that accounts for the information in the measurements. In Section 4 we describe the signal processing we employ to extract information from the Doppler signal.

In Section 5 we describe the classification mechanism that we employ to combine the evidences from the Doppler and the audio signal. We use a simple Bayesian mechanism within which we combine the likelihoods of features derived from the Doppler and audio signals with with exponential weighting [11]. We describe experiments in Section 6 which show that this mechanism can result in significantly improved speaker ID than that obtained with audio alone. Finally in Section 7 we present our conclusions.

2. THE ACOUSTIC DOPPLER SONAR

Figure 1 shows the acoustic Doppler sodar augmented microphone that we have used in our work. It has three components. The central component is a conventional acoustic microphone. To one side of it

*Work was performed at Mitsubishi Electric Research Lab

Fig. 1. The Doppler-augmented microphone used in our experiments. The two devices taped to the sides of the central audio microphone are a high-frequency emitter and a high-frequency sensor.

is a ultra-sound emitter that emits a 40Khz tone. To the other side is a high-frequency transducer (receiver) that is tuned to capture signals around 40Khz. The microphone and transmitter are well-aligned, and placed directly pointed to the mouth. Both the emitter and receiver have a diameter that is approximately equal to the wavelength of the emitted 40kHz tone, and thus have a beamwidth of about 60^o, making them quite directional. Signals emitted by the 40Khz transmitter are reflected by the fact and captured by the receiver. It must be noted that the receiver also captures high-frequency harmonics from the speech and any background noise; however these are significantly attenuated with respect to the level of the reflected Doppler signal in most standard operating conditions. The cost of the entire setup shown in the Figure is not significantly greater than that of the acoustic microphone itself: the high-frequency transmitter and receiver both cost less than a dollar. The transmission and capture of the Doppler signal can be performed concurrently with that of the acoustic signal by a standard stereo sound card. Since the high-frequency transducer is highly tuned and has a bandwidth of only about 4Khz, the principle of band-pass sampling may be applied, and the signal need not be sampled at more than 12Khz (although in our experiments we have sampled the signal at 96Khz).

3. DOPPLER EFFECT ON SIGNALS REFLECTED FROM A TALKER'S FACE

The Doppler sonar operates on the Doppler's effect, whereby the frequency perceived by a listener who is in motion relative to the signal emitter is different from that emitted by the source. Specifically if the source emits a frequency f that is reflected by an object moving with velocity v with respect to the transmitter, then the reflected signal sensed at the emitter $\hat{f}$ is given by

$$\hat{f} = \frac{v_s + v}{v_s - v} f \tag{1}$$

were v_s is the velocity of the sound in the medium. If the signal is reflected by multiple objects moving at different velocities then multiple frequencies will be sensed at the receiver.

The human face is an articulated object with multiple components capable of moving at different velocities. When a person speaks the articulators including but not limited to the lips, tongue, jaw cheeks etc. move with velocities that depend on facial construction and are typical of the speaker. The ultrasonic signal reflected off the face of a subject has multiple frequencies each associated with one of the moving components. This reflected signal can be mathematically modeled as

$$d(t) = \sum_{i=1}^{N} a_i(t) cos(2\pi f_i(t) + \phi_i) + \Psi_{speaker} \tag{2}$$

where f_i is the frequency of the reflected signal from the i^{th} articulator, which is dependent on v_i velocity of the component. f_c is the transmitted ultrasonic frequency. $a_i(t)$ is a time-varying reflection coefficient that is related to the distance of the articulator from the sensor. ϕ_i is an articulator-specific phase correction term. The term within the summation in Equation 2 represents the sum of a number of frequency modulated signals, where the modulating signals $f_i(t)$ are the velocity functions of the articulators. We do not, however, attempt to resolve the individual velocity functions via demodulation. The quantity $\Psi_{speaker}$ is a speaker specific term that accounts for the baseline reflection from the speaker's face. It represents a crude zeroth order characterization of the bumps and valleys in the face and is not related to motion. Figure 2 shows a typical Doppler signal captured by the receiver on our Doppler sensor. The overall characteristics of this signal may be assumed to be typical of the speaker.

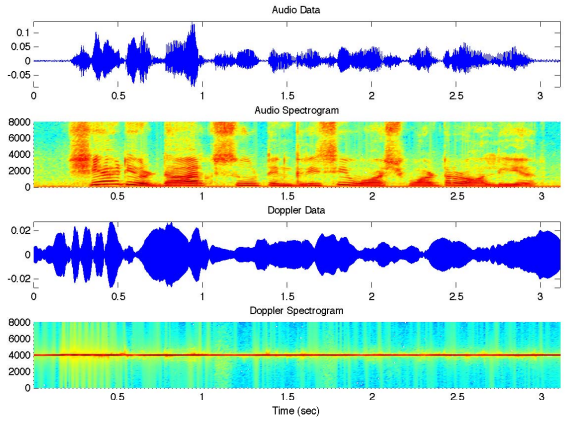

Fig. 2. A speech signal and its spectrogram, and the corresponding Doppler signal and its spectrogram.

4. SIGNAL PROCESSING

Two separate signals are captured by the Doppler augmented microphone – the microphone in the center captures the speech signal, whereas the ultrasonic transducer captures the Doppler signal. Both signals are originally sampled at 96 kHz in stereo. Since the ultrasonic sensor is highly frequency selective, the effective bandwidth of the Doppler signal is less than 8kHz, centered at 40 kHz. We therefore heterodyne the signal from the Doppler channel down by 36 kHz so that the signal is now centered at 4 kHz. Both the audio and Doppler signals are then resampled to 16 kHz.

Different signal processing schemes are applied on the Doppler and speech signals. We describe these below:

4.1. Doppler

The frequency characteristics of the Doppler signal vary slowly, since the articulators that modulate its frequency are relatively slow-moving. To capture the frequency characteristics of the Doppler signal we therefore segment it into relatively long analysis frames of 40 ms. Adjacent frames overlap by 75%, such that 100 such frames are obtained every second. Each frame is Hamming windowed and a 1024-point Fourier transform performed on it to obtain a 513-point power spectral vector. The power spectrum is logarithmically compressed

and a Discrete Cosine Transform (DCT) is applied to it. The first 40 DCT coefficents are retained to obtain a 40-dimensional cepstral vector. Each cepstral vector is then augmented by a difference vector as follows:

$$\begin{aligned} \Delta C^d[n] &= C^d[n+2] - C^d[n-2] \\ c^d[n] &= [C^d[n]^T \Delta C^d[n]^T]^T \end{aligned} \quad (3)$$

where $C^d[n]$ represents the cepstral vector of the n^{th} analysis frame, $\Delta C^d[n]$ is the corresponding difference vector and $c^d[n]$ is the augmented 80-dimensional cepstral vector. The augmented vectors are finally used for classification.

4.2. Audio

The speech signal is parameterized similarly to the Doppler signal, with the exception of the size of the analysis frames. The signal is segmented into frames of 20 ms. Adjacent frames overlap by 10ms, resulting in 100 analysis frames per second. The window shifts have been chosen to have frame-wise synchrony between the Doppler and audio channels in our setup; however this is not essential. The frames are Hamming windowed and analyzed by a 512-point FFT to obtain a 257 point power spectrum. Although it is conventional in speech recognition to integrate the power spectrum down to a Mel-frequency spectrum, we did not obtain any significant advantage from the process in the work reported here. In the experiments reported later in this paper the power spectrum was not integrated into a Mel-frequency spectrum. The power spectrum is logarithmically compressed and a DCT computed from it to obtain a 40-dimensional cepstral vector. The cepstral vector is augmented by a difference vector that is computed as the component-wise difference of the cepstral vectors from immediately adjacent frames as:

$$\begin{aligned} \Delta C^a[n] &= C^a[n+1] - C^a[n-1] \\ c^a[n] &= [C^a[n]^T \Delta C^a[n]^T]^T \end{aligned} \quad (4)$$

5. CLASSIFIERS

We use a simple Bayesian formulation for speaker identification. For each speaker, we learn a separate distribution for the feature vectors from each of the two channels (Doppler and speech). For the purpose of modeling these distributions, we assume that the sequence of feature vectors from any channel to be IID. Specifically, we assume that the distribution of both speech and Doppler feature vectors for any speaker w is a Gaussian mixture of the form:

$$P(A|w) = \sum_i c^a_{w,i} \mathcal{N}(A; \mu^a_{w,i}, R^a_{w,i}) \quad (5)$$

$$P(D|w) = \sum_i c^d_{w,i} \mathcal{N}(D; \mu^d_{w,i}, R^d_{w,i}) \quad (6)$$

where A and D represent a random feature vectors derived from speech and Doppler signals respectively. $P(A|w)$ and $P(D|w)$ represent the distribution of speech and Doppler feature vectors for speaker w, respectively. $\mathcal{N}(X; \mu, R)$ represents the value of a multivariate Gaussian with mean μ and covariance R at a point X; $\mu^a_{w,i}, R^a_{w,i}$ and $c^a_{w,i}$ represent the mean, covariance matrix and mixture weight respectively of the i^{th} Gaussian in the distribution of speech feature vectors for speaker w, while $\mu^d_{w,i}, R^d_{w,i}$ and $c^d_{w,i}$ represent the mean, covariance matrix and mixture weights for the i^{th} Gaussian in the distribution of Doppler features for the speaker. All parameters of all distributions are learned from a small amount of joint Doppler+speech recordings from the speaker.

Classification is performed using a simple Bayesian classifier. Let $\{\mathbf{A}\ \mathbf{D}\}$ represent the set of all speech and Doppler feature vectors obtained from any test recording of a subject. The subject is recognized as a speaker $\hat{w}$ according to the rule:

$$\hat{w} = \text{argmax}_w P(w) \prod_{A,D \in \mathbf{A}\ \mathbf{D}} P(A|w)^\alpha P(D|w)^{1-\alpha} \quad (7)$$

where $P(w)$ represents the *a priori* probability of the speaker w. We assume the probability to be uniform for all the subjects. α is a positive weight term that lies between 0 and 1.0 and represents the confidence we have in the likelihood obtained from the audio measurements. It can be estimated from a held-out test set. More typically, α must be varied with the background noise level: increasing noise can affect the speech signal (or Doppler signal, if the noise has very high frequencies and is energetic enough to be captured by the ultrasonic sensor), and consequently, α must be varied to increase reliance on the Doppler signal as the relative dependability of the speech signal reduces.

Figure 3 shows the block diagram of the overall Speaker Identification system that combines evidence from both the Doppler and speech channels.

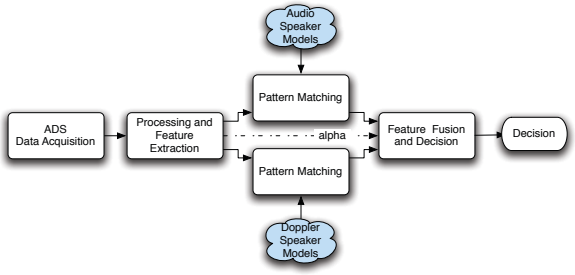

Fig. 3. Overall scheme for combining evidences from speech and Doppler measurements for speaker ID.

6. EXPERIMENTS AND RESULTS

Experiments were conducted to evaluate the usefulness of the Doppler signal as a secondary source of information for speaker ID. All experiments were conducted on a corpus of joint Doppler and audio recordings collected at Mitsubishi Electric Research Labs. A total of 50 native and non-native speakers were made to record 75 sentences each from the TIMIT corpus. Recordings were obtained in a sound-proofed room. All data from a speaker were recorded in a single session, although speakers were allowed to take breaks to avoid vocal fatigue. The Doppler-augmented microphone was adjusted prior to each session to point directly at the speaker's face.

The recorded data for each speaker were divided into two sets, a training set of 37 utterances and a test set of 38 utterances. Gaussian mixture densities comprising 4 Gaussians were trained for both the Doppler and Speech from each speaker. Increasing the number of Gaussians further was not observed to result in improvements on this set.

A number of experiments were conducted to evaluate speaker ID performance. In a first "base" experiment speaker ID tests were

run on the clean test recordings for each speaker. Since the Doppler measurements are "secondary" measurements that are not affected by audio noise (particularly since the sensitivity of the Doppler sensor to far-field noise is low), they may also be expected to improve speaker ID performance under noisy conditions. To test this hypothesis additional experiments were conducted after corrupting the speech channels (only) of the test data with babble and white noise to 0db and 10dB SNR. In each case α, the parameter that governs the relative contribution of speech and Doppler to the classification was varied. Table 1 shows the results obtained in each case for different values of α.

Table 1. Speaker Recognition Accuracy

α	Clean Speech	Babble		White	
		0 dB	10 dB	0 dB	10 dB
0	81.63	81.63	81.63	81.63	81.63
0.2	99.41	51.36	80.53	51.21	72.45
0.5	99.63	13.59	58.49	19.10	40.41
0.8	99.34	1.84	32.99	8.30	15.36
1	99.19	0.22	17.34	3.97	7.71

Surprisingly, we observe that speaker ID performance using just the Doppler signal ($\alpha = 0$) is quite high, at 81.63%. On clean speech, while the speaker ID performance with speech alone is quite high, augmenting the speech signal with the Doppler at $\alpha = 0.5$ results in further improvement, reducing the error by 54% relative to that obtained with speech alone.

The addition of any noise at all to the speech results in dramatic reduction of performance of speech-only speaker identification. In all cases we are simply better off depending only on the Doppler data for speaker ID ($\alpha = 0$). The results are however, not surprising since the Doppler signal itself was not corrupted. Nevertheless, considering the relative insensitivity of the Doppler sensor to noise, it may be expected that in real-life noisy scenarios, the use of secondary Doppler information could improve speaker ID performance significantly.

7. DISCUSSION

We note overall that the Doppler sonar is an effective secondary sensor that can effectively augment speech signals for greatly improved speaker identification. The type of information captured by the sonar is fundamentally different from that in the speech signal itself. Consequently, it is able to augment the speech signal and improve speaker ID performance even in clean conditions. Under noisy conditions, the Doppler information may be expected to be of even higher value.

Further, we speculate that in combination with a camera it might result in greater improvements still. Although the Doppler sonar captures features related to the talker's physiognomy just as a camera does, the features captured by it are fundamentally different. A camera captures a series of static images, and any image represents a snapshot of instantaneous *pose*. The velocities of various parts of the face must be arrived at by differentiation. The Doppler sensor, on the other hand captures instantaneous *velocities*. Thus, the Doppler measurements are orthogonal to those obtained by the camera, and the Doppler sensor may, in fact, be complimentary to the camera.

The Doppler sensor also does not have some of the problems associated with cameras: since the sensor is active (the ultrasonic is being beamed on the face), it does not require external signal sources, unlike cameras that cannot work in the dark. Since we capture the movements of the entire face, and reflections from objects farther from the face are typically very attenuated, there is no need to explicitly extract face-related components from the signal. Feature extraction is also not a problem – the simple technique described in Section 4 suffices. Since the emitter and receiver are collocated some of the registration/lighting issues associated with cameras do not occur.

The Doppler sensor nevertheless is susceptible both to reflections from clutter and other generators of signals in the frequency range it operates on. As part of future work, we will be addressing the issue of eliminating clutter from the signal. Further, variations in the angle of the speakers face affect measurements – the sensed velocities depend on the angle of the face. We expect that we can normalize out some of these variations at least through the use of adaptive transformations of the captured spectra, and through the use of multiple ultrasonic receivers. These and other issues will the topics of future research.

8. REFERENCES

[1] C. che and Q. Lin, "Speech recognition using hmm with experiments on yoho database," in *Eurospeech*, December 1995, pp. 625–28.

[2] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Mag*, vol. 11, pp. 1832, December 1994.

[3] A. Mezghani and D. O'Shaughnessy, "Speaker verification using a new representation based on a combination of mfcc and formants," *Canadian Conference on Electrical and Computer Engineering*, , no. 1461-1464, 2005.

[4] kajarekar and et. al, "Speaker recognition using prosodic and lexical features," in *ASRU 03*, 2003, pp. 19–24.

[5] H.B.D. Sorensen and U. Hartmann, "Pi-sigma and hidden control based self-structuring models for text-independent speaker recognition," ICASSP-93, vol. 1, pp. 537–540.

[6] J. Navratil, Qin Jin, W.D. Andrews, and J.P. Campbell, "Phonetic speaker recognition using maximum-likelihood binary-decision tree models," in *ICASSP*, 2003, vol. 4, pp. 796–799.

[7] Tieyan Fu, Xiao Xing Liu, Lu Hong Liang, Xiaobo Pi, and A.V. Nefian, "Audio-visual speaker identification using coupled hidden markov models," in *ICIP*, 2003, vol. 3, pp. 29–32.

[8] H. E Cetingul, , Y. Yemez, E. Erzin, and A.M. Tekalp, "Discriminative analysis of lip motion features for speaker identification and speech-reading," *Image Processing*, vol. 15, no. 10, pp. 2879–2891, 2006.

[9] K. Kalgaonkar, R. Bhiksha, and R. Hu, "Ultrasonic doppler for voice activity detection," *Signal Processing Letters*, vol. 14, no. 10, pp. 754–757, 2007.

[10] Bo Zhu, Timothy J. Hazen, and James R. Glass, "Multimodal speech recognition with ultrasonic sensors," in *Eurospeech*, 07.

[11] P. Beyerlein, "Discriminative model combination," in *ICASSP*, 1998, vol. 1, p. 481.