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

Synthesizing Speech from Doppler Signals

Arthur Toth, Bhiksha Raj, Kaustubh Kalgaonkar, Tony Ezzat

TR2010-016 April 2010

Abstract

It has long been considered a desirable goal to be able to construct an intelligible speech signal merely by observing the talker in the act of speaking. Past methods at performing this have been based on camera-based observations of the talker's face, combined with statistical methods that infer the speech signal from the facial motion captured by the camera. Other methods have included synthesis of speech from measurements taken by electro-myelo graphs and other devices that are tethered to the talker - an undesirable setup. In this paper we present a new device for synthesizing speech from characterizations of facial motion associated with speech - a Doppler sonar. Facial movement is characterized through Doppler frequency shifts in a tone that is incident on the talker's face. These frequency shifts are used to infer the underlying speech signal. The setup is farfield and untethered, with the sonar acting from the distance of a regular desktop microphone. Preliminary experimental evaluations show that the mechanism is very promising - we are able to synthesize reasonable speech signals, comparable to those obtained from tethered devices such as EMGs.

ICASSP 2010

SYNTHESIZING SPEECH FROM DOPPLER SIGNALS

Arthur R. Toth
Carnegie Mellon University
atoth@cs.cmu.edu

Bhiksha Raj
Carnegie Mellon University
bhiksha@cs.cmu.edu

Kaustubh Kalgaonkar
Georgia Institute of Technology
kaustubh@ece.gatech.edu

Tony Ezzat
Mitsubishi Electric Research Labratories
ezzat@merl.com

ABSTRACT

It has long been considered a desirable goal to be able to construct an intelligible speech signal merely by *observing* the talker in the act of speaking. Past methods at performing this have been based on camera-based observations of the talker's face, combined with statistical methods that infer the speech signal from the facial motion captured by the camera. Other methods have included synthesis of speech from measurements taken by electro-myelo graphs and other devices that are tethered to the talker – an undesirable setup. In this paper we present a new device for synthesizing speech from characterizations of facial motion associated with speech – a Doppler sonar. Facial movement is characterized through Doppler frequency shifts in a tone that is incident on the talker's face. These frequency shifts are used to infer the underlying speech signal. The setup is farfield and untethered, with the sonar acting from the distance of a regular desktop microphone. Preliminary experimental evaluations show that the mechanism is very promising – we are able to synthesize reasonable speech signals, comparable to those obtained from tethered devices such as EMGs.

Index Terms— Speech synthesis

1. INTRODUCTION

It has long been considered a desirable goal to be able to construct an intelligible speech signal merely by *observing* the talker in the act of speaking. Commonly, the act of observing a talker has been interpreted as one of capturing images of the talker's face. The video may then be decoded into a speech signal [1]. More commonly, observation of speech has been performed with *tethered* devices, such as electro-myelo graphs (EMG) [2], electromagnetic articulographs (EMA), or even sensors to detect brain activity. All of these signals have been demonstrated to carry sufficient information to generate speech. Regardless, devices that require tethering of wires or sensors to a talker's person are unlikely to be considered desirable. Other sensors such as Radars that capture articulator dynamics [3] have also been proposed as observation mechanisms; however the information they attempt to capture is highly specific, namely articulator configurations, and the setup is restrictive on the speaker and can become very expensive.

In this paper we propose to use a completely different device to capture movements of a talker's face, that could then be converted to speech: an acoustic Doppler sensor (ADS). The ADS is an inexpensive far-field sensor that can obtain measurements of movements of a talker's face. The device consists of a rather simple setup including an ultrasound emitter and an ultrasound sensor that is tuned to the transmitted frequency. An ultrasound tone output by the emitter is reflected from the talker's face. Movements of the talker's face impart Doppler frequency shifts to the reflected signal. The reflected "Doppler" signal now contains a spectrum of frequencies that represent the motion of the speaker's cheeks, lips, tongue, jaw, etc. The pattern of movements of facial muscles is indicative of the sound generated and may be used to infer the actual speech signal. Ultrasound Doppler signatures from the ADS have previously been used successfully for voice activity detection [4], speaker identification [5] and even speech recognition [6, 7]. Here we show that they can be used to synthesize speech as well.

The Doppler sensor is different in nature from other non-tethered "observational" sensors used in this context in the past. Cameras, radar-like devices and ultra-sound sensors have typically been used to capture *positional* information, either in the form of images or in the form of reflection delays for pulsed ultrasound bursts. The ADS on the other hand captures *velocity* information. The Doppler frequency shifts in the signal represent velocities of facial components. The spectrum of the Doppler signal may hence be viewed as a *velocity spectrum* of the face.

Our basic approach is to use techniques from Voice Transformation (VT) to convert the doppler data to features that can be used to synthesize speech. The original goal of VT was to map speech from one person's identity to another's [8]. Viewed more generally, VT can be seen as a technique to map correlated features to features that can be used to synthesize

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.

speech. This approach has been used to synthesize speech from Electro Magnetic Articulograph (EMA) data [9], Non-Audible Murmur (NAM) data [10], and surface Electrmyography (EMG) [2].

It is not clear that the VT method is sufficient to derive information from the Doppler signal. The frequency shifts resulting from facial movements are proportional to the velocity of facial features and are very small, often below the resolution of a regular DFT. Nevertheless, our experiments show that we are able to derive sufficient information from the Doppler signals using conventional spectral analysis to be able to synthesize reasonable speech signals out of the Doppler measurements, albeit thus far only in a speaker-specific manner.

2. THE ACOUSTIC DOPPLER SENSOR

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 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 emitter 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 face and captured by the receiver. It must be noted that the receiver also captures 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, not counting the microphone, is less than $20 using off-the-shelf components. The microphone was included in our setup in order to be able to record training data required by the technique outlined in Section 4. 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 receiver 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. In practice we sampled both audio and Doppler at 96kHz. The Doppler signals were heterodyned down by 36kHZ, such that the carrier frequency is shifted to 4kHz, and both speech and ultrasound were resampled at 16 kHz.

3. THE DOPPLER SIGNAL

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 $\hat{f} = (v_s + v)/(v_s - v)f$, where 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. Figure 2 shows a typical Doppler signal captured by the receiver on our Doppler sensor. The speech, the corresponding Doppler signal, and the spectrograms of both are all shown. We note that the spectral information in the Doppler signal is has very narrow spread and is hard to resolve. This is what we must obtain cues from to synthesize speech.

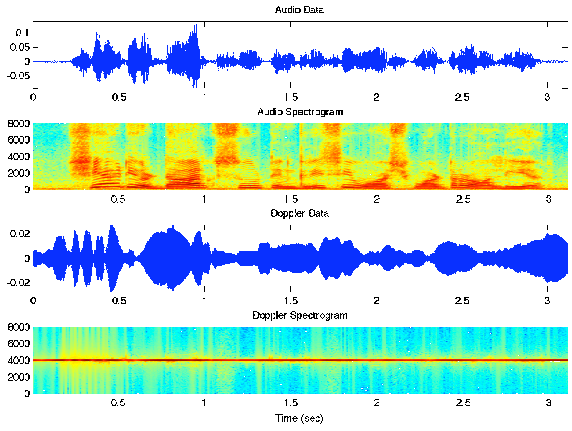

Fig. 2. A speech signal and its spectrogram, and the corresponding Doppler signal and its spectrogram. For synthesis from Doppler signals the spectrogram of panel 2 must be estimated from that in panel 4.

4. SYNTHESIZING SPEECH FROM DOPPLER SIGNALS

The transformation system used to obtain speech from Doppler signals is based on modified versions of voice-transformation tools that are freely available from the FestVox project [11]. The general overall procedure is as follows.

A training corpus of joint Doppler and speech recordings is recorded using the Doppler-augmented microphone. The Doppler signals (by which we refer to the down-heterodyned, down-shifted and down-sampled ultrasound signal) are perfectly synchronized with the speech signal. The speech signals are parameterized into a sequence of 24-dimensional mel-frequency cepstral coefficients. These didn't not include the zeroth cepstral feature which represents the energy in the analysis frame. The Doppler signals are also similarly parameterized into 25-dimensional cepstral coefficients. These included the energy zeroth cepstral coefficient. Analysis windows were 25ms, with a 15ms overlap between adjacent windows, for both signals. The cepstral features were augmented with difference features for both signals. This results in 48-dimensional extended features for the speech and 50-dimensional features for the Doppler. Finally, the resulting extended feature vectors obtained from the Doppler and Speech signals were concatenated to result in a single 98-dimensional feature vector for each analysis frame.

A Gaussian-mixture distribution was then trained from the collection of 98-dimensional feature vectors derived from the training corpus.

During synthesis, when only the Doppler signal is available and the speech signal must be inferred, 50-dimensional cepstral+difference features were derived from the Doppler signal. These were used to obtain maximum *a posteriori* estimates of the corresponding 48-dimensional features for the speech using the GMM learned in the training phase. The procedure described in [12] was used to derive the sequence of spectral envelopes for the speech signal.

The spectral envelopes must be scaled by the estimated energy of the speech in each analysis frame, and must be excited using the appropriate pitch – F_0 (or by noise for unvoiced frames). In principle, both of these can be predicted using the same procedure that is used to predict the rest of the cepstral features for the speech. For this paper, however, we tried two variants. In the first, we derived the power and F_0 values directly from an auio signal captured jointly with the Doppler signal. This is not a practical strategy for synthesizing speech from doppler data, because it assumes you already have the speech, but it is useful way to isolate the spectral conversion map and assess its quality.

In the second type of transformation we only used the 0th through 24th doppler MCEPs as input features. They were used both for the prediction of audio spectral features and audio power. Following the lead of previous researchers [10] we avoided the difficulties of F_0 prediction by treating transformed utterances as completely unvoiced. This strategy transformed doppler data to whisper-like speech. This approach is valid for synthesis from doppler as no audio data is used during transformation, however, the best way of handling F_0 estimation remains open to investigation.

N	MCD mean (std.dev.)
1	7.24 (1.97)
2	7.25 (2.10)
4	7.05 (2.04)
8	6.84 (1.98)
16	6.69 (1.92)

Table 1. *Means and standard deviations of the MCD between the actual speech signal and the signal synthesized from corresponding Doppler recordings*

5. EXPERIMENTS

We conducted experiments to evaluate the synthesis of speech signals from Doppler measurements. A set of 188 sentences from TIMIT were read by a subject into the Doppler-augmented microphone. The subject was ask to generally face the mic, but no other impositions were made in order to ensure that the speech and facial motions were natural. Simultaneous Doppler and audio recordings were made, with the final signal captured as 16kHz as mentioned in Section II.

In our first experiment we investigated the ability of doppler features to predict speech spectral features. For this experiment, fundamental frequency and power estimates for the synthesized speech were derived from the audio signal. The system was trained on 170 utterances from the and testing was performed on 18 utterances. The results of these trials according to the Mel-Cepstral Distortion (MCD) metric are in Table 1. The "N" column lists the number of Gaussians in the mixture model. The "MCD mean (std. dev.)" column lists the means and standard deviations of the MCDs between the estimated audio MCEPs and the actual MCEPs extracted from the audio data. MCD is a scaled Euclidean distance and is a popular metric used in voice transformation. Smaller numbers are better as they represent a closer match between the speech synthesized from the Doppler and the actual speech signal observed. The MCDs from these experiments are on the high side of those seen during typical voice transformations from speech to speech, but are not outside the range of what has been seen.

Informal listening indicated that parts of the synthesized utterances could be understood, and other parts, though unintelligible, clearly sounded like speech. Spectrograms of a recorded speech signal, and the signal synthesized from the corresponding Doppler are shown in Figures 3 and 4, respectively. Visual inspection suggests that a large portion of formant structure is predicted by the doppler-to-speech mapping.

In the second experiment we only used the Doppler signal itself to synthesize speech, with no energy or F_0 cues

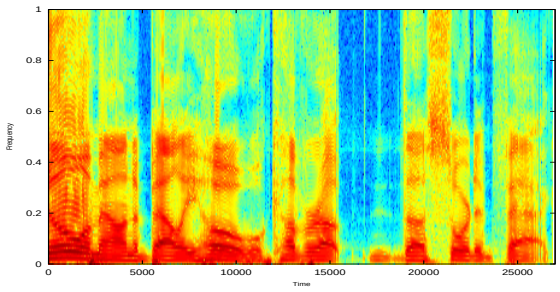

Fig. 3. Audio Spectrogram for Recorded Utterance: "No amount of ballyhoo will cover up the sordid fact."

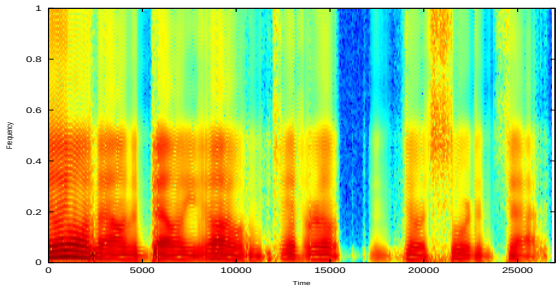

Fig. 4. Audio spectrogram for synthesized utterance where spectral features were predicted from the Doppler signal, but power and F_0 were taken from the audio recording.

derived from the audio. The spectrogram for the synthesized utterance of this type for the same utterance as the other spectrograms is in Figure 5. Again, visual inspection suggests that a good deal of format structure is predicted. One thing to note is that treating the utterance as unvoiced leads to using noise excitation during synthesis, and this causes the resulting spectrogram to differ from the previous ones in that the "ripples" associated with the F_0 curve and its harmonics are missing. Examples of synthesized audio (both whispered, and using F_0 from the speech) may be heard at `http://www.cs.cmu.edu/~bhiksha/audio/doppler`

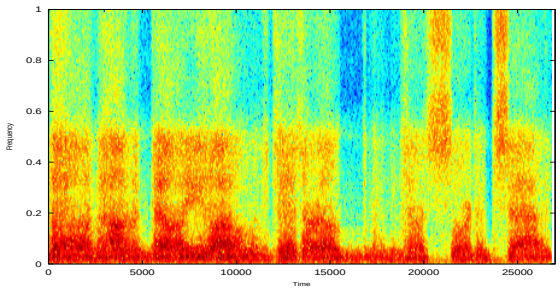

Fig. 5. Audio spectrogram for a whispered signal synthesized entirely from Doppler measurements.

6. CONCLUSIONS

Our experiments reveal that it is possible to derive at least partially intelligible speech from the Doppler signal. We believe that the synthesis can be significantly better. In this paper we have only used a very crude cepstral characterization of the Doppler signal. However, as we note from Figure 2, the spectral information is the Doppler signal is hard to resolve with conventional spectral processing. Better characterizations are possible using longer analysis windows and through modulation spectra that characterize instantaneous frequency, for example. The statistical model for the Doppler-to-speech mapping is a simple Gaussian mixture model. Doppler information is *dynamic* by nature, representing velocities rather than position. We believe that better synthesis may be obtained using more complex models such as switching linear dynamic models. The experiments conducted in this paper were not truly *silent* – the subjects actually spoke. We also propose to extend our method to situations where the speaker merely mimes speech. All of these are topics for future investigation.

7. REFERENCES

[1] T. Hueber, Benaroya E.-L., Chollet G., Denby B., Dreyfus G., and Stone M., "Visuo-phonetic decoding using multi-stream and context-dependent models for an ultrasound-based silent speech interface," in *Interspeech*, 2009.

[2] A. Toth, M. Wand, and T. Schultz, "Synthesizing speech from electromyography using voice transformation techniques," in *Interspeech*, 2009.

[3] Holzrichter J., "Characterizing silent and pseudo-silent speech using radar-like sensors," in *Interspeech*, 2009.

[4] Kalgaonkar K., Raj B., and Hu R., "Ultrasonic doppler for voice activity detection," *IEEE Signal Processing Letters*, vol. 14(10), pp. 754–757, 2007.

[5] Kalgaonkar K. and Raj B., "Ultrasonic doppler sensor for speaker recognition," in *ICASSP*, 2008.

[6] Srinivasan S., Raj B., and Ezzat T., "Ultrasonic sensing for robust speech recognition," in *Submitted to ICASSP*, 2010.

[7] Zhu B., Hazen T. J., and Glass J. R., "Multimodal speech recognition with ultrasonic sensors," in *Eurospeech*, 2007.

[8] D. G. Childers, B. Yegnanarayana, and Ke Wu, "Voice conversion: Factors responsible for quality," in *ICASSP*, 1985.

[9] T. Toda, A. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis," in *5th ISCA Speech Synthesis Workshop*, June 2004.

[10] M. Nakagiri, T. Toda, H. Kashioka, and K. Shikano, "Improving body transmitted unvoiced speech with statistical voice conversion," in *Interspeech*, 2006.

[11] A. Black and K. Lenzo, "Building voices in the Festival speech synthesis system," http://festvox.org/bsv/, 2000.

[12] T. Toda, A. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *ICASSP*, 2005.