Audio Privacy: Reducing Speech Intelligibility while Preserving Environmental Sounds

Francine Chen
FX Palo Alto Laboratory
3400 Hillview Ave, Bldg 4
Palo Alto, CA USA
chen@fxpal.com

John Adcock
FX Palo Alto Laboratory
3400 Hillview Ave, Bldg 4
Palo Alto, CA USA
adcock@fxpal.com

Shruti Krishnagiri
FX Palo Alto Laboratory
3400 Hillview Ave, Bldg 4
Palo Alto, CA USA
krishnagiri@fxpal.com

ABSTRACT

Audio monitoring has many applications but also raises privacy concerns. In an attempt to help alleviate these concerns, we have developed a method for reducing the intelligibility of speech while preserving intonation and the ability to recognize most environmental sounds. The method is based on identifying vocalic regions and replacing the vocal tract transfer function of these regions with the transfer function from prerecorded vowels, where the identity of the replacement vowel is independent of the identity of the spoken syllable. The audio signal is then re-synthesized using the original pitch and energy, but with the modified vocal tract transfer function. We performed an intelligibility study which showed that environmental sounds remained recognizable but speech intelligibility can be dramatically reduced to a 7% word recognition rate.

Categories and Subject Descriptors

H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing—*Signal analysis, synthesis, and processing*

General Terms

Algorithms, Measurement, Performance, Human Factors

Keywords

audio privacy, remote awareness, surveillance

1. INTRODUCTION

Audio communication can be an important component of electronically mediated environments, such as virtual environments and remote collaboration systems, and remote monitoring situations, such as security surveillance, home monitoring of the elderly, or always-on remote awareness. However, privacy concerns are often raised in connection with remote monitoring (e.g., [1]). In addition to providing a verbal communication channel, audio can also provide useful contextual information without intelligible speech. By rendering the speech unintelligible, privacy concerns may be mitigated while preserving this useful contextual information.

If environmental sounds are preserved when processing to reduce speech intelligibility, monitoring can be performed to identify sounds of interest. By preserving prosodic information, that is, pitch and relative energy, a listener should be able to tell if someone sounds distressed. In the security scenario, sounds such as glass breaking, gunshots, or yelling are indicative of events that should be investigated. In the elder care scenario, examples of sounds which might indicate intervention is needed are a tea kettle whistling for a long time, the sound of something falling, or the sound of someone crying. By preserving the nature and identifiability of environmental sounds and prosodic information, the audio monitoring system can provide valuable remote awareness without overly compromising the privacy of the monitored. Such a monitoring system could be valuable in augmenting a system with the ability to automatically detect important sounds, since the list of important sounds can be diverse and possibly open-ended. Such a system could also be used as an alternative to or an extension of video monitoring.

In this paper, we describe an automatic method for reducing the intelligibility of speech while preserving prosody and non-speech environmental sounds. We evaluated our approach with an intelligibility study showing that our approach significantly reduces the intelligibility of speech while allowing for identification of other sounds.

2. RELATED WORK

Schmandt and Vallejo [8] present one approach to preserving audio privacy in a monitoring scenario. They performed explicit detection of events which are then represented with fixed audio cues/icons. When speech is detected, it is rendered unintelligible with a time-based randomization process. It does not try to preserve environmental sounds that may occur simultaneously with speech. Because the system only detects sounds that are prespecified and will ignore others, unexpected important sounds may be missed or mis-represented.

A different approach is to distort the audio in a way that impacts only speech intelligibility. This eliminates the need for identifying a set of important sounds and implementing classifiers. Although it is trivial to process speech to render it unintelligible, preserving prosody and environmental sounds at the same time is much more difficult.

In the speech perception literature, Cole et al. [3] and Kewley-Port et al. [6] conducted studies on the amount of information carried by different types of speech sounds. Their work indicates that most of the information needed to recognize words is carried in the vowels and very little is carried in the consonants.

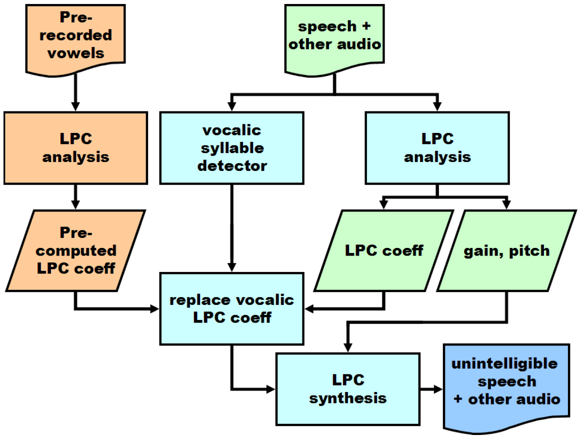

Figure 1: System Overview

Our work draws on the observations in [3] and [6] that replacing vowels with noise significantly reduces speech intelligibility. In contrast to their work, we present an *automatic* method for replacing vocalics that preserves the perception of speech as well as the intonation patterns and that preserves the recognizability of processed environmental sounds.

3. SELECTIVE AUDIO OBFUSCATION

To render only speech unintelligible while preserving prosody and environmental sounds, we first process the audio signal to separate the prosodic information from the vocal tract information using Linear Predictive Coding (LPC) [7]. We then replace the LPC coefficients representing the vocal tract transfer function of the vocalics in the input speech with the stored LPC coefficients from vocalics spoken by previously recorded speakers.

By replacing only vocalic regions of speech, a sample of background sounds is heard, and we hypothesize, most of the recognizability of environmental sounds is retained. To preserve the “speechiness” of the audio, we replace the vocalic portion of a syllable with unrelated vowels produced by a different vocal tract, while retaining the speaker’s prosodic information. We also use more than one speaker to provide a further confounding effect, as [5] noted that intelligibility is better when listening to one speaker than when tested on multiple speakers.

Figure 1 is an overview of our system. On the left side of the figure, the LPC coefficients of prerecorded vowels are computed and stored as shown in orange. The input audio containing speech to be rendered unintelligible is shown in green. Voiced regions are identified in the input speech and then syllables, if any, are found within each voiced region. Given the pitch and voicing ratio computed by the voicing detector, together with the syllable detector, vocalic syllables with a pitch within the range of human speech are identified. The LPC coefficients of the identified vocalic syllables are then replaced with one of the precomputed LPC coefficients. The LPC coefficients are left unchanged for the portions of the signal that are not recognized as vocalic syllables. Using the gain and pitch computed from the original input speech together with the modified LPC coefficients, the unintelligible speech is synthesized.

3.1 Vocalic Syllable Detection

We first identify voiced segments and then identify syllable boundaries within each voiced segment. To identify voiced segments, the autocorrelation is computed every 125 ms using a tapered window of 50 ms. The pitch is estimated from the offset of the peak value of the autocorrelation function, and the ratio of the peak value of the autocorrelation to the total energy in the analysis frame provides the measure of the degree of voicing (voicing ratio). If the estimated pitch is within plausible values for adult speech and the voicing ratio is greater than a given threshold (0.2), then the speech is identified as vocalic.

Syllable boundaries are identified based on energy contour. We use the gain, G, computed from the LPC model. G is smoothed using a lowpass filter with a cutoff frequency of 100 Hz. Within a voiced segment the local minima of the energy are located and used as syllable boundaries.

3.2 Vocalic Selection

We explored using several single vocalic or combinations of vocalic sounds to replace the vocalic portion of syllables. Although most sounds significantly reduced intelligibility, most also significantly reduced the recognition of environmental sounds, introducing significant perceived distortions. Using /wa/ was hypothesized to sound like Charlie Brown’s parents. It was found to reduce intelligibility, but also to produce a distracting ”beating” sound, reducing the recognizability of environmental sounds. Use of a more neutral /ae/ sound resulted in reduced intelligibility, and most environmental sounds were still recognizable, but a sizable percentage of words were still intelligible, based on informally listening to the processed sentences.

Motivated by the reduced intelligibility noted in [5], we tried using two different replacement vowels, from two different speakers. Use of an /iy/ from a female speaker and an /uw/ from a male speaker again resulted in reduced intelligibility. However, /iy/ and /uw/ have very different vocal tract configurations, leading to an unnatural sound at the transition between two adjacent vocalic syllables. To reduce this effect, we alternated the replacement sound for vocalic regions with a male speaking /uw/ and a female speaking /uw/. This was found to reduce the unnatural transitions and reduce intelligibility while preserving environmental sounds, based on informal listening. The unnatural transitions could also be reduced in other ways, such as spectral smoothing [2]. For our experiments, we chose to alternate one male and one female exemplar of /uw/ for replacing vocalic syllables.

3.3 Vocalic Substitution

We used a 16 pole LPC model, computing pitch, gain and coefficients representing the vocal tract. Before processing any audio, the LPC coefficients, LPC_s, are computed for the selected ”substitute” vowels and stored.

To obfuscate an audio waveform the LPC coefficients of the waveform are computed and the syllables identified. For the vocalic portion of each syllable, a stored vowel is selected. The LPC coefficients representing the L frames of the selected stored vowel, $LPC_s(0, ..., L-1)$, are substituted into the LPC model for the M frame vocalic portion of identified syllable, $LPC_m(0, ..., M-1)$, replacing the first $\min(L, M)$ LPC frames. If $M > L$, then the coefficients from the last frame, $LPC_s(L-1)$ are padded until there are M frames.

Using the modified LPC coefficients, speech is synthesized with the LPC pitch and gain information computed from the original speaker, producing mostly unintelligible speech. For

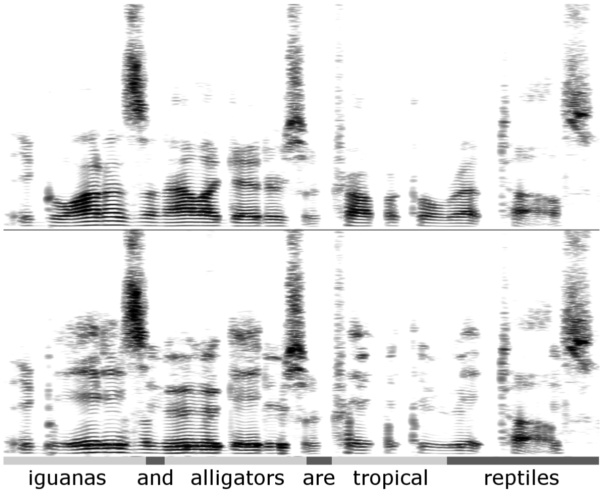

Figure 2: Spectrograms of the original (top), and processed (bottom) processed using the LPC coefficients from two other speakers saying /uw/. Vertical axis is frequency, horizontal axis is time, and darker colors denote higher magnitude. The words of the spoken phrase are roughly aligned in segments along the bottom.

most non-speech sounds, little, if any, of the sound should be identified as a vocalic syllable, and therefore, non-speech sounds in non-vocalic regions are modified only by the distortion caused by LPC modeling.

4. EVALUATION

Figure 2 shows original (top) and processed (bottom) spectrograms of the TIMIT [4] sentence denoted DR5-MDWA0-SX95. Note that the vocalic segments (strong, horizontally banded regions) in the processed version are different from the original on top, while the spectral characteristics of the non-vocalic segments, such as the final consonants of *'iguanas'* and *'alligators'*, are preserved. Note also that the pitch contour, evident in the periodic variation of the spectra, is preserved.

4.1 Intelligibility and Identifiability

We performed an intelligibility study with 12 listeners to compare the intelligibility of processed and unprocessed speech and the recognition of processed and unprocessed environmental sounds. In the study, audio files were played to listeners who were asked to distinguish the type of the stimulus (speech, sound, or both) and to identify the words and sounds they heard.

4.1.1 Stimuli

The audio stimuli were composed of 40 spoken sentences, 30 environmental sounds, and 12 containing one spoken sentence mixed with one environmental sound. The 52 spoken sentences were drawn from the TIMIT corpus [4]. To reduce the difficulty of the transcription, chosen sentences contained 5 or 6 words. The 52 sentences contained 287 total words. The sentences were drawn from 50 talkers; 2 talkers contributed two sentences with the other 48 talkers contributing just one each. Regional dialects were represented in roughly the same proportion as in TIMIT.

Fifty environmental stimuli were downloaded from several websites, including http://www.grsites.com/sounds, http://-freesound.iua.upf.edu, and http://www.wavsource.com. The 50 sounds were categorized as animal, human, home/appliances, nature/water, violence, vehicles, and miscellaneous. The peak energy of each audio file was normalized to the same value. When combining the spoken and environmental sounds, the two normalized audio files were combined and renormalized. Processing to reduce intelligibility was done on the normalized audio files.

	recognition rate (%)			
	one listen		many listens	
category	unproc	proc	unproc	proc
speech (v>95%)	98.0	6.5	100.0	17.3
env sounds	85.0	78.3	85.6	82.8
both (speech)	96.7	3.3	100.0	16.7
both (sounds)	79.2	70.8	83.3	79.2

Table 1: Recognition rates. Speech-only word % reported for stimuli with at least 95% correct voicing.

Each type of audio stimuli (speech only, non-speech, both) was randomly divided into two subsets. Each listener was presented with one subset after processing and the other subset without processing. Unprocessed audio was presented first to half of the listeners and second to the other half. The presentation order of the stimuli was also balanced across subjects to offset learning effects.

4.1.2 Study Procedure

The processed stimuli were presented in one contiguous session, and the unprocessed stimuli were presented in a second session. Before each session, a listener heard training examples of stimuli corresponding to the nature of the ensuing test. For each training session, two speech, two environmental, and two mixed sounds were used.

The 8 listeners were all adults native speakers of American English with no known hearing problems. They were asked to adjust the volume at the beginning of the session while playing a given sample of speech. The stimuli were presented through a Logitech Premium Headset 350 to both ears.

After playing a stimulus, listeners were asked to indicate whether they heard: 1) speech only 2) sound only or 3) both. If they heard any speech, they were asked to type in the words they heard. If they heard an environmental sound, they were asked to select a sound from a set of 50 sounds. “None of the above” was an option in case a listener thought an environmental sound did not sound like any of the listed sounds. Listeners were encouraged to guess if they were unsure about speech or sounds.

For each stimulus we recorded the listener response after a single presentation (to simulate a real-time monitoring scenario) and again after the listener was allowed to replay the sound as many times as desired.

4.1.3 Results

Table 1 compares word and sound recognition rates for the various conditions. The recognition rates when the stimuli was heard just once (one listen) are presented separately from when the stimuli was played as many times as desired (many listens). When computing word recognition rates the 20 most frequent stop words (e.g., 'the', 'a', 'and') were discounted. Very close matches (*e.g.*: iguanas for iguana) and misspellings were counted as correct. In Table 1, the speech only condition was evaluated over the six sentences for which the voicing detector correctly detected at least 95% of the vocalic regions in a processed sentence. The % voicing correct was computed by comparing the detected vocalic re-

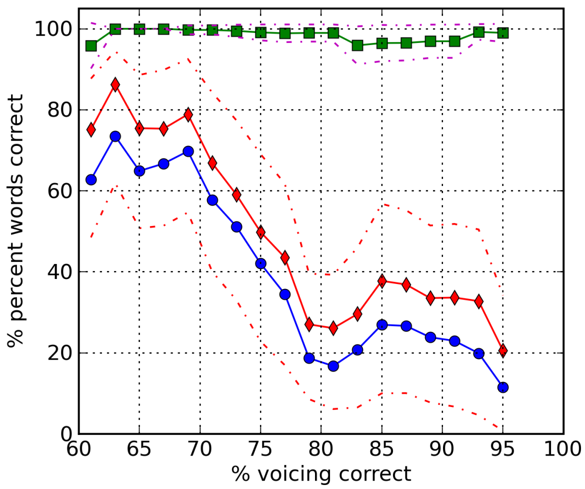

Figure 3: Word recognition accuracy vs voicing detector accuracy. Speech-only stimuli were pooled across 10% wide windows of voicing error. Performance on: original sentences, □; 1^{st} listen of processed sentences, ○; and unlimited listens, ◇. The dashed lines indicate a $\pm\sigma/2$ range about the observed means. (The upper $\pm\sigma/2$ range is omitted for 1^{st} listen and the lower for multiple listens.)

gions against regions that were hand-labeled as a vowel or sonorant in the TIMIT corpus. For speech alone, the word recognition rate is 6.5% on first listen (as in live monitoring), or approximately one word in two sentences is correctly recognized. Precision, that is, the percentage of guessed words that were correct, was 18.4%, or 4 out of 5 guessed words were incorrect. When a sentence was heard as many times as desired (as in reviewing recorded audio), the recognition rate and precision increased to 17.6% and 31.1%, respectively.

Note in Table 1 that recognition of processed environmental sounds is within seven percentage points of unprocessed audio, and much better than speech intelligibility. When speech and an environmental sound were both present, the recognition of words and environmental sounds are both a bit lower, and environmental sound recognition is still much better than speech intelligibility.

Our method is based around voicing detection, so we examined how intelligibility is influenced by errors in voicing estimation. We computed the accuracy of the voicing detector for each sentence from the TIMIT ground truth and plot the trend with a window of 10 percentage points (note that we used 5 percentage points in Table 1) in correct voicing detection as shown in Figure 3. The percentage of words correctly recognized increases when the % voicing correct dips below 80 (i.e. a range of 75-85). This result highlights the importance of the voicing detector accurately detecting voiced regions. We also examined performance when stop words are included, and the performance was very similar when comparing the trend with and without stop words.

We note that although pitch is preserved by the processing, the talker characteristics are modified because the substituted vocal tract functions used are not that of the original speaker. We also informally noted that the prosodic information is preserved and listeners can discern whether a statement or question was spoken.

5. SUMMARY AND FUTURE DIRECTIONS

We have presented an automatic method for providing audio privacy by obfuscating speech while preserving the identifiability of environmental sounds and prosodic information. Our evaluation shows that replacing vocalics with other vocalics in speech can significantly reduce speech intelligibility while preserving the recognition of most environmental sounds and the recognition of speech vs non-speech vs both speech and non-speech.

We note that the performance of the voicing detector strongly influences recognition, particularly when the performance is below 80%. While the voicing ratio is what we used to identify vocalic segments in our implementation, there are other approaches to voiced-speech identification which may be more accurate.

Other modifications to the selection of precomputed replacement vocalic segments can be performed to further decrease intelligibility of speech. These include the use of more speakers and additional replacement sounds.

In situations where it is desirable to preserve the identity of the speaker, or at least to enhance the ability to distinguish different speakers, the replacement LPC coefficients may be chosen in a speaker-dependent way based on measured parameters of the currently observed speech (mean pitch, mean spectra or cepstra, or other features useful for distinguishing talkers). And if it were desirable to further disguise the speaker, then the pitch and energy contours could also be modified. For instance the pitch could be modified by adding a random offset to each voiced segment.

6. REFERENCES

[1] K. Caine. Privacy perceptions of visual sensing devices: Effects of users' ability and type of sensing device. Master's thesis, Georgia Institute of Technology, 2006.

[2] D. T. Chappell and J. H. L. Hansen. Spectral smoothing for concatenative speech synthesis. In *International Conference on Spoken Language Processing*, volume 5, pages 1935–1938, 1998.

[3] R. A. Cole, Y. Yan, B. Mak, M. Fanty, and T. Bailey. The contribution of consonants versus vowels to word recognition in fluent speech. In *Proc. ICASSP '96*, pages 853–856, Atlanta, GA, 1996.

[4] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fixcus, D. S. Pallet, N. L. Dahlgren, and V. Zue. Timit acoustic-phonetic continuous speech corpus. Linguistic Data Consortium, Philadelphia.

[5] I. Gauthier, A. C.-N. Wong, W. G. Hayward, and O. S. Cheung. Font tuning associated with expertise in letter perception. *Perception*, 35:541–559, 2006.

[6] D. Kewley-Port, T. Z. Burkle, and J. H. Lee. Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 122(4):2365–2375, Oct. 2007.

[7] L. Rabiner and R. Schafer. *Digital Processing of Speech Signals*, chapter 7. Prentice-Hall, Inc., 1978.

[8] C. Schmandt and G. Vallejo. "listenin" to domestic environments from remote locations. In *Proc. the 2003 International Conference on Auditory Display*, pages 853–856, Boston, MA, 2003.