THE EFFECT OF POLARITY INVERSION OF SPEECH ON HUMAN PERCEPTION AND DATA HIDING AS AN APPLICATION

S. Sakaguchi, T. Arai, Y. Murahara

Dept. of Electrical and Electronics Eng., Sophia University
7–1 Kioi-cho, Chiyoda-ku, Tokyo, JAPAN

ABSTRACT

In this paper we investigate how polarity inversion of speech signals effects human perception, and we apply this technique for data hiding. In most languages, glottal airflow during phonation is uni-directional, causing constant polarity of the speech waveform. On the other hand, the human auditory system cannot discriminate between speech signals with positive and negative polarity. Based on these facts, we developed an algorithm to hide data in speech signals. We assigned one bit to each syllable of speech, and inverted the polarity of the signal at every syllable according to the assigned bit. We performed a test using 20 sentences from the TIMIT corpus to determine both whether a human could distinguish between the original and polarity-inverted signal and whether we could automatically restore the embedded binary data. We found that we were able to successfully hide data and restore it automatically.

1. INTRODUCTION

In nearly all speech sounds, the basic source of power is the respiratory system which expels air from lungs. From the pulmonary cavity, air travels through the trachea and into the larynx, at which point it passes between the vocal folds. If the folds are apart, as is normal during expiration, the air will have a relatively free passage into the pharynx [1]. Speech is usually produced during exhalation [2]. Thus, the glottal airflow during phonation is uni-directional, and it causes a constant polarity of speech waveforms.

One way to determine polarity is to estimate the glottal airflow of a speech signal by means of inverse filtering, such as with linear predictive analysis (LPC analysis) [3]. Speech sounds consist of voiced and unvoiced sounds, and they can be modeled as the response of the vocal-tract filter systems to one or more sound sources [4]. In LPC analysis, voiced sounds are assumed to be quasi-periodic pulses, and unvoiced sounds, random noise. Because the LPC residual signal for voiced sounds is impulsive, one may detect the direction of, say, vocalic airflow by observing the direction of pulses in the LPC residual signal rather than by looking at the original speech waveform.

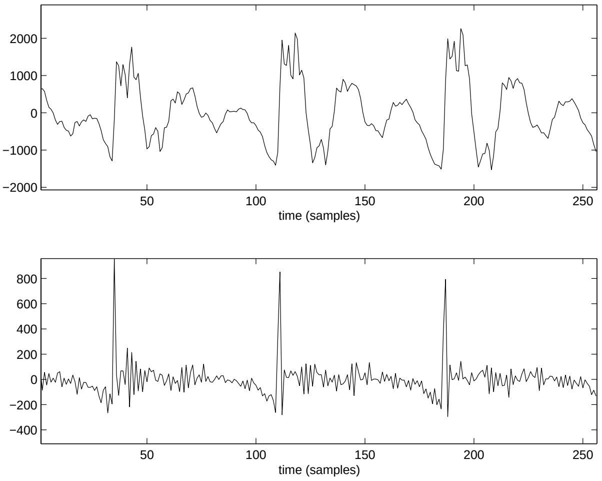

Figure 1: Speech signal (top) and LPC residual (bottom).

Figure 1 shows a typical example of the LPC residual signal of a vowel. We can find the polarity of the original speech waveform by looking at the direction of pulses of the LPC residual signal, which is upward in this case. The polarity is constant despite changes in vowel quality as long as various conditions are unchanged (i.e., speech is always produced on an exhale, the polarity of electric equipment is unchanged in time, etc.).

The human auditory system is insensitive to the phase of a simple sound, although it is sensitive to frequency and amplitude [4]. Thus, phase is often ignored in the representation of the spectrum of a signal. However, for complex sounds, such as speech, there can be a different vocal quality when two sounds differ with respect to phase alone [5]. Because the shape of the envelope is unchanged by polarity inversion of the entire signal, we expect there to be no effect on human perception whether the polarity is inverted or not.

Insensitivity of human perception to the polarity of speech suggests the possibility of data hiding. Auditory masking has already been reported for data hiding (e.g., [6]). In this method, a watermark is generated by filtering a PN-sequence with a filter that approximates the frequency masking characteristics of the human auditory system.

In this paper, we develop an algorithm to hide data in speech signals based on the facts described above, i.e., constant polarity in the speech waveform, and insensitivity of humans to polarity inversion. Because the nucleus of the syllable is a vowel, and the amplitude is comparatively small in the onset and coda, we assign one bit to each syllable of speech, and invert the polarity of the signal at every syllable according to the assigned bit. First, we manually detect syllable boundaries and conduct a perceptual experiment to investigate human perception of polarity inversion. Second, we test algorithms for automatic detection of syllable boundaries and restoration of the embedded information. We describe the principle of this study in Section 2 and our human discrimination experiment in Section 3. In Section 4 we first describe an automatic procedure for detecting syllable boundaries. We then describe conducting a human discrimination experiment between the automatically processed and the original speech, and automatically restoring the hidden data.

2. PRINCIPLE

For hiding data into a speech signal by using the polarity inversion, we divide the speech signal into segments in time. Each segment of the speech signal is then assigned one bit, and the polarity of each segment is inverted according to the assigned bit.

In determining the size of the segment, we found that the following conditions yield the best performance: 1) the amplitude is small at both ends of the segment, and 2) the segment is sufficiently long. 1) is needed to minimize the discontinuities resulting from polarity inversion. Ideally polarity would be inverted during silence or the closure portion of a voiceless stop. Inverting during consonants is also acceptable because of their low amplitude relative to the vowel nucleus. There is a trade-off in 2). The segment must be long enough to estimate polarity, but as short as possible to maximize the bit rate of data to be hidden. If we choose the sentence rather than the syllable as the unit of division, we can safely hide one bit per unit, having enough vowels to estimate the global polarity, but the bit rate would be too low. The syllable appears to be the most efficient unit of division.

A typical syllable consists of a vowel nucleus, and at the onset and the coda, the power of the waveform is usually small, and it satisfies the condition 1), above. The syllable is an appropriate length for estimating the polarity within the steady-state portion of a vowel. The average duration of syllables and vowels in American English is about 200-250 ms [7] and 100 ms [8], respectively.

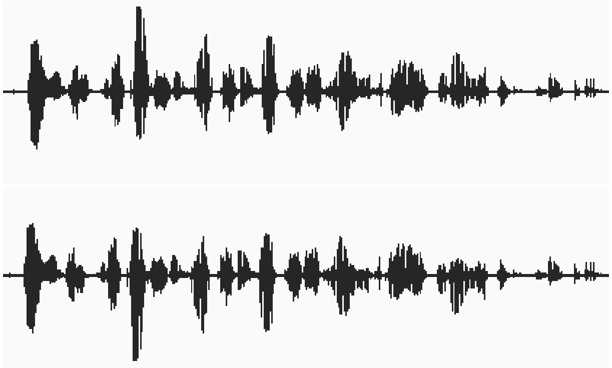

Figure 2: Speech waveforms before (top) and after (down) the polarity inversion.

Thus, each syllable (or syllable-like unit) is assigned one bit, and its polarity is inverted at every syllable according to the assigned bit. To extract the embedded bit stream, the polarity of the LPC residual is detected.

3. POLARITY INVERSION EXPERIMENT

We conducted a human perceptual experiment using the original and polarity-inverted speech. Twenty sentences in TIMIT corpus were used as speech samples, and the syllable boundaries were detected manually.

3.1. Detecting syllable boundaries

We divided the speech signal into segments containing several syllables each. Segments were broken up during nonvocalic portions of the waveform. Where the amplitude was small, certain segment boundaries were determined manually.

3.2. Inverting the polarity

Each syllable was assigned one bit, and we inverted its polarity according to the assigned bit. To minimize the effect of discontinuity at the boundary of each segment, we applied a trapezoidal window, linearly tapering 8 ms of both ends. For the perceptual experiment, we embedded the binary sequence ‘1 0 1 0 1 0 · · ·’. That is, the polarity was inverted alternately at each syllable. Figure 2 shows the original and the processed speech waveform.

3.3. Human discrimination experiment

It is essential that the quality of the sound not become distorted in the process of data hiding. We, therefore, conducted a discrimination experiment to determine whether humans could discriminate between the original and polarity-inverted speech signals. Twenty TIMIT sentences were used for the experiment (16-kHz sampling and 16-bit quantization). The subjects were 20 native speakers of Japanese, and no subjects reported having any previous hearing problem. An ABX discrimination test was used for this experiment, where X is either A or B. When A is the original signal, B will be the processed signal, and vice versa. The order of the stimuli was randomized for a different subject. The experiments were conducted in a sound proof chamber, using a PC. The subject used a headset (Sennheiser, HD 600) to listen to the stimuli, and followed instructions on the PC display. Subjects were able to listen to the same stimulus up to 10 times.

3.4. Experimental results

The average rate of correct answers in this experiment was 51.3% both when X was the original and the processed signal. This result tells us that humans did not discriminate between the processed and original speech signals.

4. AUTOMATION OF PROCESSES

Manually detecting syllable boundaries and restoring binary data by hand is time consuming, and the results are inconsistent. Therefore, we implemented an algorithm to detect syllable boundaries automatically. We then conducted an experiment to see whether humans could distinguish the original from the polarity-inverted signal obtained by using the automatically detected boundaries. We also implemented an algorithm to automatically restore the hidden data.

4.1. Automatic detection of syllable boundaries

In this section we describe an algorithm which automatically detects syllable boundaries. Wu et al. [9] proposed an algorithm for automatic detection of syllable onset based on subband energy contours. In the present study, we use a combination of the full-band energy contour and voicing detection using modified autocorrelation, in order to ensure that there are voiced segments (vowels) between two consecutive boundaries. We used the same 20 TIMIT sentences as were used in the previous experiment. The speech signal was truncated by a Hamming window; the frame was 256-point long (16 ms) and 50% overlapped. We then obtained an LPC residual signal where the order was 16. Next, we calculated autocorrelation of the LPC residual signal for each frame. We judge the frame to be voiced when the modified autocorrelation had a peak with a normalized peak value of 0.2 and a peak lag was greater than 2.5 ms. Finally, the boundary was determined to be the point of minimum energy contour within each portion of consecutive possibly unvoiced frames. A boundary was rejected if it falls less than 200 milliseconds after the previous one.

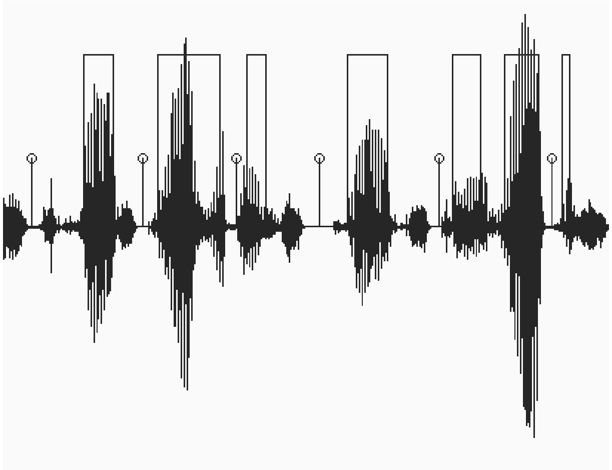

Figure 3: Estimated vowel locations (rectangular wave) and syllable boundaries (unfilled circles) obtained by the proposed algorithm.

Figure 3 shows an example of syllable-like boundaries determined with our algorithm.

4.2. Another human discrimination experiment

Based on the boundaries obtained by the automatic detection algorithm described in Section 4.1, we inverted the polarities of the 20 TIMIT sentences according to the same procedure described in Section 3.2. We conducted another discrimination experiment like that described in Section 3.3, this time using automatically located syllable boundaries. The rate of correct answers in this experiment was 53.0% on the average. From this result, it was confirmed that humans cannot discriminate between the original and polarity-inverted speech when using automatically detected syllable boundaries.

Some specific sentences had a slightly higher rate of correct answers. By looking at the data closely, we found that this is because the automatic algorithm detected vowel transitions as boundaries and as a result the polarity inversion caused slight discontinuities in the signal.

4.3. Automatic extraction of hidden information

We tried to extract hidden information automatically from the processed speech signals. Our automatic algorithm is based on LPC analysis. First, syllable boundaries are detected automatically as in Section 4.1. Then, the polarity of the maximum peak within the LPC residual signal of each voicing frame is estimated. Finally, we determine the polarity of the syllable (or syllable-like unit) by taking the most commonly estimated polarity from among voicing frames within the syllable.

We made several binary streams and inverted the polarity of each syllable according to the assigned bit for 20 sentences. The average number of bits per sentence was 7.6. We then extracted the embedded bit streams from the processed speech signals and compared the extracted bit stream with the original. The rate of correct bits was 96.7% (there were five errors out of 152 bits). There were two types of errors: those based on the automatic syllable-boundary detection (two errors) and errors based on the automatic bit extraction (three errors).

5. CONCLUSION

We confirmed that human listeners are insensitive to the polarity of syllables of speech signals. And we successfully hid data in speech signals by inverting the polarity of syllables, a process which was undetectable to human listeners. Although the bit rate of information that we can hide is not that high, this technique is extensible. For example, one could perfect the automation algorithm and develop a technique for making watermarks which would hide such things as copyright information by considering the robustness against attacks and various signal manipulations.

6. ACKNOWLEDGMENTS

We gratefully acknowledge the members of Arai Laboratory of Sophia University who gave us useful comments, as well as the participants in the perceptual experiments. We would like to express our grateful thanks to Terri Lander.

7. REFERENCES

[1] P. Ladefoged, *A Course in Phonetics* (3rd Ed.), Harcourt Brace Jovanovich College, 1993.

[2] D. O'Shaughnessy, *Speech Communication*, Addison-Wesley, 1990.

[3] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, Springer-Verlag, 1980.

[4] G. Fant, *Acoustic Theory of Speech Production*, Mouton, 1960.

[5] R. C. Mathes and R. L. Miller, "Phase Effects in Monaural Perception," *J. Acoust. Soc. Am.*, vol. 19, pp. 780–797, 1947.

[6] L. Boney, A. H. Tewfik and K. N. Hamdy, "Digital Watermarks for Audio Signals," *Proc. MULTIMEDIA* (*IEEE*), pp. 473–480, 1996.

[7] S. Greenberg, "On the Origins of Speech Intelligibility in the Real World," *Proc. ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 23–32, 1997.

[8] S. Greenberg, J. Hollenback, D. Ellis, "Insights into Spoken Language Gleaned from Phonetic Transcription of the Switchboard Corpus," *Proc. ICSLP*, pp. S32–35, 1996.

[9] S. Wu, M. L. Shire, S. Greenberg and N. Morgan, "Integrating Syllable Boundary Information into Speech Recognition," *Proc. ICASSP* (*IEEE*), 1997.