



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Real-Time Speech Masking Using EM-Wave Acoustic Sensors

J. F. Holzrichter, L. C. Ng, J. T. Chang

November 28, 2012

38th International Conference on Acoustics, Speech, and
Signal Processing (ICASSP)

Vancouver, Canada

May 26, 2013 through May 31, 2013

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

REAL-TIME SPEECH MASKING USING EM-WAVE ACOUSTIC SENSORS

Holzrichter J.F.*; Ng L.C.[#]; and Chang J.T.*

*Lawrence Livermore National Laboratory, PO Box 808, L-20, Livermore, CA 94550, [#]LLNL retired

ABSTRACT

Micro-power EM-wave sensors positioned on or near a speaker's Adam's apple have advantages over other acoustic techniques for enabling speech sound reduction and masking. For voiced speech, they provide glottal closure information with 0.1 ms timing accuracy (i.e., ~10kHz bandwidth), and provide closure timing up to ~0.5 ms in advance of the corresponding acoustic speech signal leaving the speaker's mouth. Unvoiced and silent speech segments are readily identified by the absence of vocal fold motion, by their acoustic signatures, and by timing. These unusual data acquisition characteristics enable anti-speech acoustic waves and prior recorded acoustic waves to be generated, then synchronized, and re-broadcast for purposes of speech masking with much improved timing and fidelity compared to prior approaches.

Index Terms -- radar, speech processing, speech masking, sensors

I. INTRODUCTION AND PRIOR WORK

The technology of "speech masking", as used in this paper, includes purposeful processes by which a human speaker's acoustic speech signal, propagating in air, is made less detectable and/or less understandable to another human listener. For this paper, we also include speech intensity reduction (i.e., cancellation) in this category. It is well known that the signal extraction and processing power of the human ear-brain system is such that complete cancellation and/or obscuration of speech sounds have been virtually impossible to accomplish in realistic situations. However, useful reductions in the perception of speakers' voices and of inanimate sounds are being achieved in office environments [1] and in noisy airplane cabins [2].

For individual speakers who wish their communications to be unintelligible to a listener, prior work shows that speech signals can be reduced in intensity up to about -10db and partially masked [3] [4]. A microphone records the speech, a processor inverts it, and then transmits it a few meters "down stream" (faster than the speech wave). There the anti-speech signal is rebroadcast "just in time" to cancel the user's speech signal at the location of a listener. This concept works best in open spaces with few acoustic reflections and where the loudspeaker's radiation pattern at the point of a listener coincides with the user's acoustic radiation pattern. Masking on the other hand, is a technique often used to make listening difficult by broadcasting various sound patterns non-synchronously, often during periods when there is a speech sound to be masked [1]. Flooding an area with white or pink background noise, music, or speakers' voices (e.g., "babble") is commonly used for simple masking.

This paper describes an improved masking process by which a human speaker uses miniature radar for measuring their vocal-fold activity [5,6]. The radar sensor accurately detects voiced speech excitation in advance of the corresponding acoustic signal leaving the speaker's lips or nose. This occurs because the acoustic sound wave takes ~0.5 ms to travel up and out of the vocal tract and into the space in front of the speaker. Miniature radar sensors for use in speech characterization such as ASR, de-noising, and other applications have been described extensively [5, 6, 7]. In addition, radar sensors also detect the absence of voiced speech, enabling techniques to be used for effectively masking unvoiced or silent speech periods. In contrast to most other acoustic voice activity detectors [8], radar-like sensors can have bandwidths of >10 kHz and are immune to acoustic noise. Hence their signal acquisition time is >10 times faster than an acoustic signal traveling through local throat tissues, and >100x faster than the time taken by a natural voiced acoustic wave traveling from the vocal folds to the nose or lips. Their unique time resolution enables the sub-millisecond timing needed to generate improved real-time cancellation and masking. The radar signals are sufficiently synchronous with the corresponding acoustic speech signal that effective amplitude reduction and phoneme modification can be accomplished. An important aspect of partially cancelled original speech is that masking signals can be added to it without increasing the original speech

amplitude as perceived by a listener. These techniques enable superior noise masking of an individual speaker's voice compared to prior work described above and elsewhere.

In particular, we show in this paper how to use the prior glottal and corresponding acoustic signal, from the just-completed speech period, about 5 ms prior to the presently being generated signal. We then show how to construct an anti-speech signal matched to the present acoustic signal as it leaves the speaker's mouth and nose. First the system tests the recorded prior period acoustic signal for continuity, timing, makes corrections as needed, and then re-broadcasts the anti-speech signal through a nearby loud-speaker. In most causes, useful cancellation of the present-period speech-signal occurs in a real world setting, when proper electronics and loudspeaker are used, and when multipath effects are minimal. Figs 2 and 3 illustrate how prior recorded radar and speech data are time and amplitude adjusted, and then inverted for masking purposes. In essence, these techniques enable the construction of partially cancelled, continuous sound level, complex "cocktail" noises that reduce a speaker's intelligibility at a distant listener.

II. EQUIPMENT AND DATA

Figure 1 illustrates the present experimental set up.

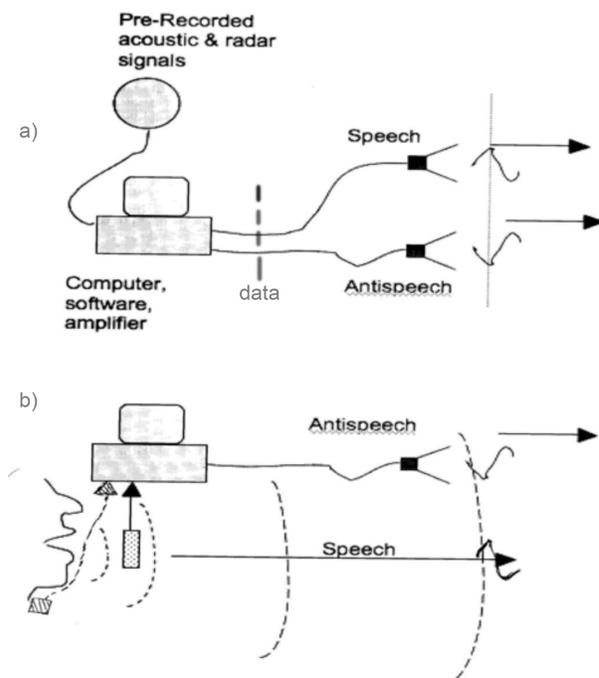


Fig 1: Illustration of experimental apparatus. a) Present experiments with prior recorded data sampled at the location of the dotted-line noted as "data". b) Illustration of a real-time application with a micro-power radar on a speaker's neck and microphone in front of mouth of a speaker. In experiments described below, anti-speech and other masking signals are generated and then broadcast synchronously (i.e., coherently) with a user's natural sound signal.

Speech experts have described [8] how voiced phoneme signal shapes change slowly from glottal period to glottal period, even when phoneme sounds change. In prior work [9] [10]., the authors showed the relationships between typical voiced speech glottal shapes and a micro-power radar sensor. The slowly changing acoustic and radar signals (compared to modern electronics speeds) are easily followed using a radar or other sensor of sufficient time bandwidth to obtain absolute and relative timing, amplitude, and other data for subsequent masking. In those cases when the well behaved glottal patterns stop or change dramatically, such as at the onset of a first or at the end of the last glottal cycle in a voiced speech segment, the radar sensor detects such changes and an algorithm tests for unvoiced or silent speech. Such short-time events contain relatively little acoustic energy and usually can be partially cancelled or masked, degrading their information content (i.e., masked).

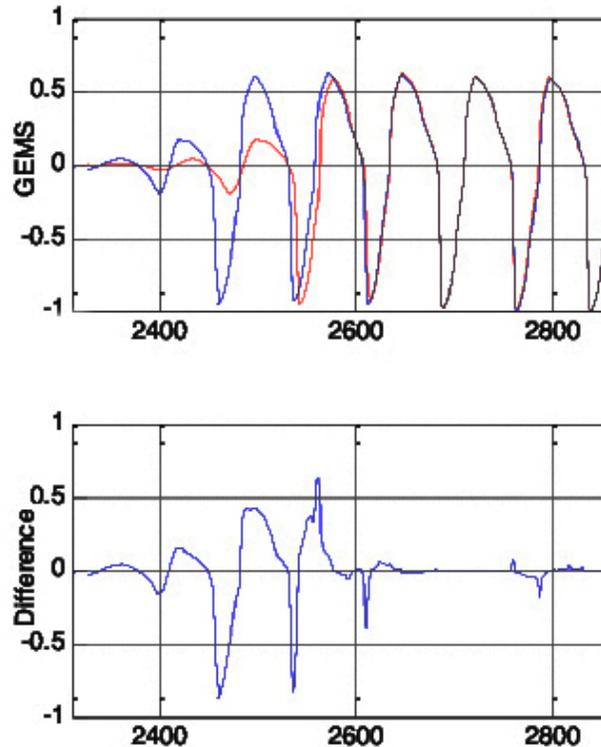


Fig. 2 Period-to-period timing study of a radar sensor signal. Time in ms is obtained by dividing horizontal axis values by 100. GEMS is a type of radar sensor [5]. Our cancellation algorithm is illustrated by shifting prior data forward in time by one voiced period to match a present radar glottal signal. It is then subtracted from the present period signal. A simple zero crossing algorithm is used for timing. The shifted radar signal is the red signal (it is the trace with lowest amplitude at time 2450, but by time 2600 it closely follows the present signal). Mismatches occur at onset or at times of rapid acoustic phoneme changes compared to the nominal 5 ms glottal cycle times. Note that after 3 periods, the algorithm correctly aligns both the radar signals (lower signal). The corrected timing is used to “time-stamp” the corresponding acoustic signal for subsequent acoustic cancellation (see Fig. 3). Most of the radar residuals arise from the incomplete cancellation at high frequencies, associated with the sharp closure of the glottis. The small residual at time 2800 illustrates the problem. Such radar residuals lead to incomplete acoustic cancellation, causing less effective acoustic speech cancellation.

III. SPEECH CANCELLATION

The term speech cancellation is used herein to mean a degree of useful sound signal reduction, in - dB. Fig.3 shows “pretty good” cancellation of an acoustic signal, which has been timed by using the corresponding radar signal in Fig 2. When using average quality acoustic sensors, a simple zero-crossing algorithm will often not work well because of their limited bandwidth and noise susceptibility. In these experiments, accurate glottal closure timing (< 0.1 ms accuracy) is needed to identify accurately the corresponding onset of a voiced acoustic time period. Each voiced period is so timed, hence when the prior speech period acoustic signal is recalled, it can be adjusted and broadcast to cancel the presently spoken acoustic signal with up to 10kHz electronic accuracy (n.b.: many other real-world elements do degrade this bandwidth).

Several human vocal system effects cause noticeable errors at the glottal-period level using the simple algorithms described. They include sequential speech signal onset miss-timing due to prosodic effects, amplitude mismatching, vocal tract shape changes as new phonemes are spoken, and others. Many of these can be accommodated by adaptive techniques, since most such changes are continuous and well behaved. For most normal speech sequences, such changes in timing and amplitude are measured to occur at the 1-2 percent per speech period rate.

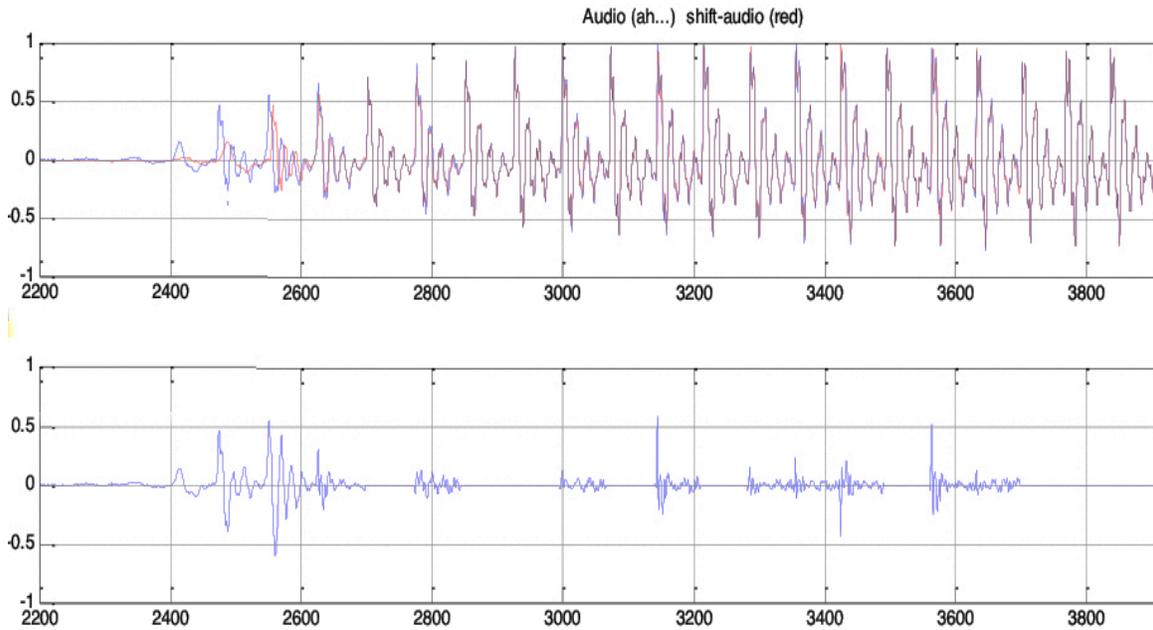


Fig. 3 : Computer experiment showing voiced acoustic speech reduction. A prior-period recorded voiced sound /ah/, is offset forward in time by one voiced period; then subtracted from the preent period. The corresponding radar signal, used for timing the voiced speech production, is shown above in Fig. 2. Residuals of the acoustic cancellation are shown in the lower panel. The larger errors (i.e., residuals) appear at onset of speech when the software takes 2 or 3 periods to optimize the timing, as well as during some sound changes. No forward adaption was employed to correct for slow changes

IV. CONCLUSION

The procedures described above enable one to contemplate much improved speech denial techniques, when compared to prior methods. In initial studies, we have shown that over 90% of a user's speech signal can be reduced in intensity and virtually the entire acoustic signal, be it voiced, unvoiced, or silence, can be "optimally" masked to reduce comprehension by a listener. This is possible because both the cancellation and masking signals are time synchronized with the user's voice production mechanisms (e.g. coherently). In particular using prior voiced-acoustic information enables a cancellation signal to be prepared during the ~0.5 ms before the presently voiced acoustic signal leaves the mouth. It is then broadcast simultaneously with the speaker's natural sound to enable a useful degree of cancellation. Other types of speech such as silent speech periods need little processing and unvoiced segments require "colored" noise or voiced speech segments for masking. To further reduce intelligibility, synchronous masking uses recorded segments of back-ground noise, prior recorded segments of the user's own speech, as well as other types of sounds which are inserted in phase with the production cycle of the user's speech being spoken.

Micro-power radars and other skin contact sensors have been considered to be too complicated for normal use, even by experts needing speech information protection, and other desirable properties. However, increasing needs for speech fidelity in noisy environments, for user verification, for improved speech recognition, and for privacy are likely to be met in the future by body-worn devices as discussed here.

V. ACKNOWLEDGEMENTS

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

REFERENCES

- [1] Hillis, Ferren, Howe, Eno , U.S. Patent No. 7,143,028 “Method and System for Masking Speech” Nov. 28, 2006 , updated 2009
- [2] See for example noise cancelling earphones from Bose Corporation
- [3] Kondo K. and Nakagawa K, "Speech emission control using active cancellation," *Speech Communication*, 49, Issue 9, September 2007, pp 687
- [4] Senders J., U.S. Patent Appl. US 20110105034 A1 ("Active Voice Cancellation System,"), May 5, 2011
- [5] One type of micro-power radar sensor was designed by T. E. McEwan at the Lawrence Livermore Laboratory (LLNL) (see patent literature). An early version was often referred to as “GEMS”, standing for Glottal Electro-Magnetic Sensor. It was used for many of the experiments described in this paper.
- [6] Holzrichter J.F., Burnett G.C., Ng L.C., and Lea W.A. “Speech articulator measurements using low power EM-wave sensors” *J. Acoust. Soc. Am.* Volume 103, Issue 1, pp. 622-625 (1998)
- [7] Ng, L.C.; Burnett, G. C.; Holzrichter, J. F.; and Gable, T.J.; “Denoising of Human Speech using Combined Acoustic and EM Sensor Signal Processing” p 229, *Proceedings 2000 IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP 2000)*
- [8] For a status of classical voice activity detection, abbreviated as “VAD” , please see the Wikipedia review of this topic. In addition, throat-skin mounted microphones, electro-glottal graph (EGG) sensors, and other devices detect onset of voiced speech with time responses in the millisecond range or faster. They may be consistent with the procedures discussed above.
- [8] Titze, I.R. “Principles of Voice Production” Prentice Hall, New Jersey,1994
- [9] Titze, I.R., Story B. H., Burnett G.C., Holzrichter J.F., Ng L.C., Lea W.A. “Comparison between Electro-glottography and Electromagnetic-glottography, *JASA* 107 pt. 1), 581 (2000)
- [10] Holzrichter J.F., Ng, L.C., Burke G.J., Champagne, N.J., Kallman J.S., Sharpe R.B., Kobler J.B., Hillman R.E., Rosowski J.J. , “Measurements of Glottal Structure Dynamics” *JASA* 117 (3), Pt.1, 1373 (2005)