Characterizing Silent and Pseudo-Silent Speech using Radar-like Sensors

John F. Holzrichter, PhD
President, Hertz Foundation
Senior Scientist, Livermore Laboratory

Sept 7, 2009
Interspeech 2009
Brighton, England

www.johnholzrichter.com
Silent and Pseudo Silent Speech with micro-Radars

- Silent Speech, related topics and micro-Radar sensors
- EM Radar-like sensors and reflected EM waves
- Silent Speech
- Pseudo-Silent Speech
- Active Cancellation & Masking
- Conclusion
Speech articulator transitions & air pressure induced motions are measured using micro-Radar sensors

Cheek, tongue, palate

e.g., Homodyne Sensor Signal

Vocal fold “machinery”

Rear section of tracheal wall
Silent & Pseudo-Silent Speech with micro-Radars

- Silent Speech in this talk means:
  - Measuring vocal tract articulator positions and/or their transitions in real time,
  - Under conditions of **Zero acoustic excitation**, 
  - Associating micro-Radar signals with meaningful symbols 
  - Example: Tongue-to-Palate contact in time ordered pattern

- Pseudo-Silent speech means:
  - Excitation of the vocal tract is allowed, 
  - External **acoustic energy is transmitted** but “cloaked”.
  - By-standers are unaware or unable to understand sound
  - Examples: murmuring, sub-vocalization, whispering, humming,
  - “Silencing” implies hiding, masking, or canceling.
EM sensor under chin measures Tongue-to-Palate contact and then the release, robustly!

Contact timing can be used for silent signaling, e.g., Morse Code
Review of Articulator measuring with micro-Radars

• EM interferometric (homodyne) sensors actively interrogate **ALL** Articulator - Air interface conditions in propagation path,

• It is easiest to measure diphone, triphone, and other transitions => **relative changes** in signal amplitude and time, which are stored as a sequence of numbers (i.e., feature vector)

• Our group’s past work concentrated on
  Characterizing the excitation and filter functions of human speech, using both radar and acoustic sensors together

• **This talk** concentrates on estimating the effectiveness of EM sensors for Silent and Pseudo Silent Speech

• Many Opportunities exist to **improve Human -> Machine** interface. higher bit rate (> 10bits/sec), more robust, hands free, etc. are enabled by using micro-Radar technologies
More about presently used micro-Radar sensors

- ~ 5 Hz to > 1000 Hz Target bandwidth (DC is usually filtered away), Range ~ 5-20 cm range inside head and neck tissue - $\lambda$ dependent

- Present longitudinal accuracy ~ 4 cm, but resolution $\Delta$~ 1 $\mu$m!

- Transverse accuracy ~ 5 cm dependant on antenna and distance

- Interface-reflected-EM waves inform Articulator states due to:
  1) Interface motion (longitudinally): e.g. tongue surface
  2) interface disappears/appears => signal change: e.g., vocal fold contact, tongue palate contact, lip contact

- Cheap RF electronics and excellent design methodologies enable small, 1-5 GHz, EM radar interferometers, with < microwatt of average power

- Quantity costs of $5 ea., sizes < 1cm, blue tooth compatible,
The human vocal tract has Excitations (e.g., vocal folds) & Articulators - They move from “state to state”
EM sensors measure vocal fold motion at low sound (air flow) levels => hum, murmur, etc.

Provides a very good excitation function

* G. Burnett thesis

Interspeech 9/09 Brighton, J.F.Holzrichter
Micro-Power Radar-like sensors measure articulator interface locations and their transitions

- Silent Speech, Related Topics and micro-Radar sensors
- EM Radar-like sensors and reflected EM waves
- Silent Speech
- Pseudo-Silent Speech
- Active Cancellation & Masking
- Conclusion
2 examples of low power radar-like sensors:
2.3 GHz pulsed interferometer & Diode Laser

Early Prototype Board version

RF IC version

LED or Laser Diode version

Interspeech 9/09 Brighton, J.F. Holzrichter
Micropower EM sensors have many applicable modalities

- EM radar-like sensor modalities that have been used:
  - Wave packet transmit with range gate
  - Chirped transmit with received signal frequency-range bins
  - Wave packet transmit and homodyne receive
    Interferometer mode
  - High pass sensing (i.e., DC “clutter” is suppressed)
  - Radar microphone

- Present experiments using homodyne sensors generate 10-ns long EM-wave packets at 2.4 GHz frequency ($\lambda = 12$ cm, but $\lambda = 1.5$ cm in tissue!), with a 2-MHz prf (typical) (1/e tissue penetration is 3 to 5 cm, one way)
Information from EM micro-Radar sensors depends upon their mode of use ~ Near Field

Near & Intermediate Field Mode

\[ r \ll \lambda < A^{1/2} \]

\[ \text{Signal} = \text{const. } \Delta \varepsilon \times A_{\text{eff}} \times (d/dr \text{ Wave}) \]

Far Field Mode (e.g., Radar)

\[ r >> \lambda > A^{1/2} \]

\[ \text{Signal} = \text{const. } A_{\text{eff}} \times 1/r^4 \]

Interspeech 9/09 Brighton, J.F.Holzrichter
Multiple EM-waves, homodyne detection, and AC filtering have been used for excitation detection.

Sensitivity function versus distance of object from sensor.

Integrator

7 kHz

High and Low pass filters

Glottal Signal

Interspeech 9/09 Brighton, J.F.Holzrichter
EM sensed tube wall motions have been measured, simulated, and understood (Holzrichter et al., JASA Mar. 05)

- Anterior Wall
  - 12-micron movement,
  - Typical 40 mV signal
  - Incident EM wave (corresponding to 400 mV signal)

- Posterior Wall
  - 12-micron movement:
  - Typical 4 mV signal

Sub-glottal pressure changes of 5-10 cm H₂O
Silent Speech information is obtained using micro-Radars to detect articulator motions

- Silent Speech, Related Topics and micro-Radar sensors
- EM Radar-like sensors and reflected EM waves
  - Silent Speech
  - Pseudo-Silent Speech
- Active Cancellation & Masking
- Conclusion
“Silent speech” requires enough articulator- state or state-transition information to define a speech unit.

- The location (i.e., “state”) of vocal structures is harder to measure - best done with ranged gated sensors

- “state transitions” are simpler since they are “self referencing” & they constrain phone sequences,

- As feature vectors are being formed, they need to be translated into a language, for the user or for a machine

- Sound feedback to user is helpful to keep him/her engaged

- “Synthetic languages” might be used, e.g, Morse code from tongue-palate contact, lip-to-lip contact, etc.
Radars under Jaw-Palate gives relative distance (silent), and vocal-fold open-close (w/ acoustics)

silence to “O N E” to silence

silence /wa/ /ah/ /nn/ silence

Jaw Drops 0.2 sec. in advance of voicing onset

Jaw - Palate Signal

Vocal Fold Signal

Time ( seconds )

Interspeech 9/09 Brighton, J.F.Holzrichter
An iPhone with EM sensor(s) can detect reflections: e.g., 3 articulator locations: lips, tongue tip, tongue hump
A typical *user-device* might interrogate a speaker’s mouth and oral cavity to obtain simple commands

- 1) Multi-Articulator Measurements:
  - 3 States of Lips: closed, open round protruding, open wide
  - 3 states of tongue: tip down, tip-palate contact, fast tip motion
  - 2 states of Jaw: open, closed
  This example gives: **18 combinations/ 0.1 sec**
    => **180 possible different feature vectors/sec**

  All 18 articulator interface permutations are measured during each round trip of EM wave. As the wave reflects off of each interface, it produces many reflections adding up to a composite reflected signal. As time progresses, stationary interfaces are removed by filtering, and differently moving interfaces, each with varying amplitudes of reflection, show-up as different signal patterns over subsequent time periods of articulator speech-state transitions. A significant number of phones can be defined, but not all.

- 2) With added Range location measurements:
  With range data added to above example, one might obtain 3 more sets of data, giving 8 more combinations of:
    2 states of tongue hump, 2 states of pharynx, and 2 states of tongue edge, for a total of **1080 potentially different feature vectors/sec**.
Pseudo Silent Speech and micro-Radar sensors

- Silent Speech, Related Topics and micro-Radar sensors
- EM Radar-like sensors and reflected EM waves
- Silent Speech
  - Pseudo-Silent Speech
  - Active Cancellation & Masking
- Conclusion
“Pseudo Silent speech” relaxes “silence” constraint, and simply demands that speech be undetectable.

- Excitation data provide a great deal more information than articulator location/transition data alone:
  1) Vocal fold motions and pressure induced vibrations
  2) “perfect” timing of each pitch period

- “State Transitions” are still easiest to measure, although filter functions for each state become easily estimated.

- Causes vocal tract to be fully engaged, especially for frication

- Enables use of full tool kit of source-filter technology, i.e., EM sensor/acoustic signal processing. (see Holzrichter, et al)

- Microphone also delivers user feedback, and makes “Cloaking” possible using < 1mW level masking and/or canceling signals.
Pseudo Silent Speech examples include:

- **Humming**: vocal folds engaged, no vocal tract changes, mouth closed.
  - Many opportunities for using fold on-off, pitch-change, and “pitch-slide” to code messages.
  - Mouth-closed leads to oral cavity excitation and cheek vibrations.
  - Easiest to hide, mask, or cancel since the sounds are “simple” and low intensity.

- **Murmuring and sub-vocalization**: vocal folds engaged,
  - Mouth often open,
  - Vocal tract engaged,
  - Low intensity => easiest to “cloak”
  - Allows use of prior acoustic-EM-sensor speech algorithms

- **Whispering**: No vocal folds => fricative excitation,
  - Vocal tract engaged, low intensity.
  - Easily hidden, masked but not cancelled (re chaotic excitation)
EM sensed organ motions enable acoustic "Cloaking", e.g., masking, canceling, hiding

- Silent Speech, Related Topics and micro-Radar sensors
- EM Radar-like sensors and reflected EM waves
- Silent Speech Information
- Pseudo-Silent Speech Information
  - Active Cancellation & Masking
- Conclusion
Hiding, Masking, and Canceling => Cloaking

- Lots of time from excitation to mouth emission, delay of \(~ 0.5\text{ms}\) from vocal-fold closure to acoustic phone emission (depends on mike placement)

- Relies on EM sensor signal to "forecast" future speech events \(\sim 0.5\text{ ms}\) in advance, e.g., silence, voiced, unvoiced

- Most speech energy is in voiced segments which are easiest to predict and "cloak".

- Fricative segments occur \(< 0.5\text{ sec}\) before or after voiced speech segments, => hide with pink noise
Cloaking: Hiding, Masking, and Canceling

*Making Pseudo Silent Speech into Silent Speech*

- Easy to measure background noise for adaptive masking

- Far Field “useful cancellation” is enabled because of small acoustic radiator-size (i.e., antenna) relative to the distance-to-listener

- Crosstalk control and feedback needed during active masking and/or canceling (as well as for silent speech)

- Requires sound feedback to speaker, so as not to confuse speaker during broadcasting of masking or canceling signals.
Diagram of “cloaked” speech

- **Masking:** detects type of speech signal and background, adds pink & warped background noise
- **Canceling:** inverts signal, adds pink & background noise
EM sensed speech organ motions enable many applications

- Silent Speech, Related Topics and Radar-like Sensors
- EM Radar-like sensors and reflected EM waves
- Silent Speech
- Pseudo-Silent Speech
- Active Cancellation & Masking
- Conclusions
Some of many ways to use micro-radars:

- Attached to skin
- Under skin, or under tape
- Neck band, or under hat, ....

Interspeech 9/09 Brighton, J.F.Holzrichter
Some other ways to use micro-radars:

- Master-of-the-Universe Communicator
- Disguised as Jewelry

Interspeech 9/09 Brighton, J.F.Holzrichter
Many applications of EM sensor/speech data are possible - Silent, Pseudo Silent, other:

- Sensor on top of head (under hat) sensor
- Cheek-oral cavity sensor
- iPhone, with Masking

Masking or canceling signal

Interspeech 9/09 Brighton, J.F. Holzrichter
Conclusion: Modern Micro-radar sensors can extract speech information with low -> “zero” acoustic signatures

- There is strong demand for minimizing noise pollution, eves-dropping, and work place distractions. E.g., airports, offices, coffee shops, police work, etc.

- Higher speed, more robust human-to-system communications are enabled by joint Acoustic-microRadar technologies

- Machine control - cameras, power tools, wheelchairs, automobiles - Work better with “super reliable” micro-Radar sensed speech

- There are several ways to minimize and perhaps eliminate acoustic signatures from hand held phone users.

- Many collateral benefits accrue:
  Waterproof, hands-free, higher baud rates, speaker identification, transmission compression, many user modalities, and more.

THANK YOU PROF. DENBY AND CONFERENCE ORGANIZERS

Interspeech 9/09 Brighton, J.F.Holzrichter