RESEARCH ARTICLE OPEN ACCESS

# Talking Without Talking

Deepak Balwani*, Honey Brijwani*, Karishma Daswani*, Somyata Rastogi*
*(Student, Department of Electronics and Telecommunication, V. E. S. Institute of Technology, Mumbai

***ABTRACT***
The Silent sound technology is a completely new technology which can prove to be a solution for those who have lost their voice but wish to speak over phone. It can be used as part of a communications system operating in silence-required or high-background- noise environments. This article outlines the history associated with the technology followed by presenting two most preferred techniques viz. Electromyography and Ultrasound SSI. The concluding Section compares and contrasts these two techniques and put forth the future prospects of this technology.
***Keywords***- Articulators , Electromyography, Ultrasound SSI, Vocoder, Linear Predictive Coding

## I. INTRODUCTION

Each one of us at some point or the other in our lives must have faced a situation of talking aloud on the cell phone in the midst of the disturbance while travelling in trains or buses or in a movie theatre. One of the technologies that can eliminate this problem is the **'Silent Sound' technology**.

'Silent sound' technology is a technique that helps one to transmit information without using vocal cords which was developed at the Karlsruhe Institute of Technology[1]. It enables speech communication to take place when an audible acoustic signal is unavailable. The main goal of 'Silent Sound' technology is to notice every movement of the lips and internally transform the electrical pulses into sounds by neglecting all the surrounding noise, which could help people who lose voices to speak, and allow people to make silent calls without bothering others. Rather than making any sounds, the handset would decipher the movements made by one's mouth by measuring muscle activity, then convert this into speech that the person on the other end of the call can hear[2].

The technology opens up a host of applications, from helping people who have lost their voice due to illness or accident to telling a trusted friend your PIN number over the phone without anyone eavesdropping — assuming no lip-readers are around. It can be used used in Military for communicating secret/confidential matters to others
Also, given the numbers of cell phones in use today, the market for this technology could potentially become very important if such a concept gained public acceptance.

Silent Sound Technology is implemented using two methods. They are

- Electromyography (EMG)
- Ultrasound SSI

Electromyography involves monitoring tiny muscular movements that occur when we speak and converting them into electrical pulses that can then be turned into speech, without a sound being uttered. Ultrasound imagery is a non-invasive and clinically safe procedure which makes possible the real-time visualization of the tongue by using an ultrasound transducer.

### 1.1. HISTORICAL FRAMEWORK

The idea of interpreting silent speech with a computer has been around for a long time, and came to public attention in the 1968 Stanley Kubrick science-fiction film ''2001 – A Space Odyssey", where a ''HAL 9000" computer was able to lip-read the conversations of astronauts who were plotting its destruction. Automatic visual lip-reading was proposed as an enhancement to speech recognition in noisy environments [3], and patents for lip-reading equipment able to interpret simple spoken commands began to be registered in the mid 1980's [4]. The first ''true" SSI system which deployed 3 electromyographic sensors mounted on speaker's face and interpreted the speech with an accuracy of 71%, originated in Japan. [5]A few years later, an imaging-based system, which extracted the tongue and lip features from the video of speaker's face, returned 91% recognition[6]. A major focal point was the DARPA Advanced Speech Encoding Program (ASE) of the early 2000's, which funded research on low bit rate speech synthesis ''with acceptable intelligibility, quality, and aural speaker recognizability in acoustically harsh environments", thus spurring developments in speech processing using a variety of mechanical and electromagnetic glottal activity sensors [7]. The first SSI research papers explicitly mentioning cellphone privacy as a goal also began to appear around this 2004[8]

## II. SILENT SOUND TECHNOLOGIES

This section illustrates two fundamental techniques that are put to use in interpretation of speech in noisy environments in the absence of intelligible acoustic signals. These techniques are:

- Electromyography
- Ultrasound SSI

Electromyography involves monitoring tiny muscular movements that occur when we speak and converting them into electrical pulses that can then be turned into speech, without a sound being uttered. Ultrasound imagery is a non-invasive and clinically safe procedure which makes possible the real-time visualization of the tongue by using an ultrasound transducer.

### 1.2. ELECTROMYOGRAPHY

Electromyography (EMG) is a technique for evaluating and recording the electrical activity produced by skeletal muscles[9]. EMG is performed using an instrument called **ELECTROMYOGRAPH,** to produce a record called an **ELECTROMYOGRAM.** An electromyograph detects the electrical potential generated by muscle cells when these cells are electrically or neurologically activated. The signals can be analysed to detect medical abnormalities, activation level, or to analyse the biomechanics of human or animal movement.

Fig 1 shows the basic mechanism of Electromyography where the muscle activity is analysed by the Electromyograph to generate an Electromyogram.

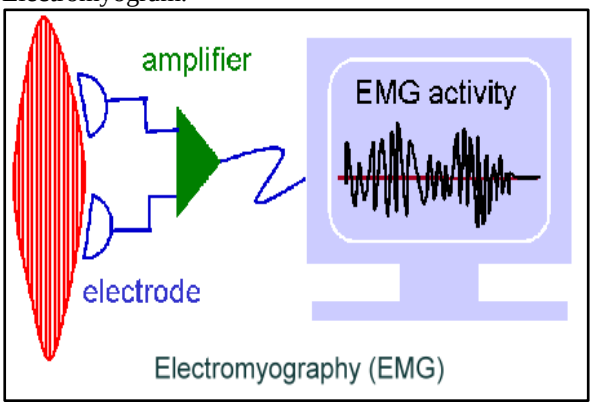


**Fig 1. ELECTRMYOGRAPHY[10]**

#### 1.2.1. ELECTRICAL CHARACTERISTICS

The electrical source is the muscle membrane potential of about -90 mV. Measured EMG potentials range between less than 50 μV and up to 20 to 30 mV, depending on the muscle under observation. Fig 2 shows EMG sensors connected on the face of a speaker.Typical repetition rate of muscle motor unit firing is about 7–20 Hz, depending on the size of the muscle. In interfacing we use four different kind of transducers -

- Vibration sensors
- Pressure sensor
- Electromagnetic sensor
- Motion sensor

**Fig 2: Electromyographic sensors attached to the face[11]**

#### 1.2.2. TYPES OF EMG

There are two kinds of EMG in widespread use: surface EMG and intramuscular (needle and fine-wire) EMG. To perform intramuscular EMG, a needle electrode or a needle containing two fine-wire electrodes is inserted through the skin into the muscle tissue. A trained professional (such as a neurologist, physiatrist, chiropractor, or physical therapist) observes the electrical activity while inserting the electrode. Certain places limit the performance of needle EMG by non-physicians. A recent case ruling in the state of New Jersey declared that it cannot be delegated to a physician's assistant. The insertional activity provides valuable information about the state of the muscle and its innervating nerve. Normal muscles at rest make certain, normal electrical signals when the needle is inserted into them. Then the electrical activity when the muscle is at rest is studied. Abnormal spontaneous activity might indicate some nerve and/or muscle damage. Then the patient is asked to contract the muscle smoothly. The shape, size, and frequency of the resulting electrical signals are judged. Then the electrode is retracted a few millimeters, and again the activity is analyzed until at least 10–20 motor units have been collected. Each electrode track gives only a very local picture of the activity of the whole

muscle. Because skeletal muscles differ in the inner structure, the electrode has to be placed at various locations to obtain an accurate study.
Intramuscular EMG may be considered too invasive or unnecessary in some cases. Instead, a surface electrode may be used to monitor the general picture of muscle activation, as opposed to the activity of only a few fibers as observed using an intramuscular EMG. This technique is used in a number of settings; for example, in the physiotherapy clinic, muscle activation is monitored using surface EMG and patients have an auditory or visual stimulus to help them know when they are activating the muscle.

The technique primarily used for this purpose is surface electromyography. Surface Electromyography (sEMG) is the process of recording electrical muscle activity captured by surface (i.e., non-implanted) electrodes. When a muscle fiber is activated by the central nervous system, small electrical currents in the form of ion flows are generated. These electrical currents move through the body tissue, whose resistance creates potential differences which can be measured between different regions on the body surface, for example on the skin. Amplified electrical signals obtained from measuring these voltages over time can be fed into electronic devices for further processing.As speech is produced by the activity of human articulatory muscles, the resulting myoelectric signal patterns measured at these muscles provides a means of recovering the speech corresponding to it. Since sEMG relies on muscle activity alone, speech can be recognized even if produced silently, i.e., without any vocal effort, and the signal furthermore cannot be corrupted or masked by ambient noise transmitted through air. As a result, sEMGbased speech recognition overcomes the major shortcomings of traditional speech recognition, namely preserving privacy of (silently) spoken conversations in public places, avoiding the disturbance of bystanders, and ensuring robust speech signal transmission in adverse environmental conditions.This technique could enable silent speech interfaces, as EMG signals are generated even when people pantomime speech without producing sound. Having effective silent speech interfaces would enable a number of compelling applications, allowing people to communicate in areas where they would not want to be overheard or where the background noise is so prevalent that they could not be heard. In order to use EMG signals in speech interfaces, however, there must be a relatively accurate method to map the signals to speech.[12]

Recent research studies aim to overcome the major limitations of today's sEMG-based speech recognition systems and applications, to, for example:

- remove the restriction of words or commands spoken in isolation and evolve toward a less limited, more user-friendly continuous speaking style
- allow for acoustic units smaller than words or phrases, enabling large vocabulary recognition systems
- implement alternative modeling schemes such as articulatory phonetic features to enhance phoneme models
- study the effects of electrode re-positioning and more robust signal preprocessing
- examine the impact of speaker dependencies on the myoelectric signal
- investigate real-life applicability, by augmenting conventional speech recognition systems and addressing size, attachment, and mobility of the capturing devices

The applicability of EMG-based speech recognition in acoustically harsh environments, such as first responder tasks where sirens, engines, and firefighters breathing apparatus may interfere with reliable communication, has been investigated at NASA. For example, Jorgensen and colleagues ( Betts et al., 2006) achieved 74% accuracy on a15-word classification task, in a real-time system which was applied to subjects exposed to a 95 dB noise level.[13]

Electromyography thus captures electrical stimuli from the articulator muscles or the larynx, which can subsequently be exploited in speech processing applications. One may also imagine, however, capturing viable speech bio signals directly from the brain, using electroencephalography (EEG) or implanted cortical electrodes. These possibilities are discussed in the following two sections. Although considerably further off in terms of commercial application, these Brain Computer Interface (BCI) approaches – very much in vogue today – are fascinating, and hold enormous promise for speech, as well as for other types of applications.

### 1.3. ULTRASOUND SSI

Another way to obtain direct information on the vocal tract configuration is via imaging techniques. Ultrasound imagery is a non-invasive and clinically safe procedure which makes possible the real-time visualization of one of the most important articulators of the speech production system – the tongue. An ultrasound transducer, placed beneath the chin, can provide a partial view of the tongue surface. Ultrasound device which is coupled with a standard optical camera as shown in Fig 3 is used to capture tongue and lip movements. Because of its non-invasive property, clinical safety and good resolution,

ultrasound is well adapted to vocal tract imaging and analysis. Furthermore, since laptops with high performance ultrasound imaging systems are available, a portable real-time SSI system with an ultrasound transducer and camera, can be feasible.

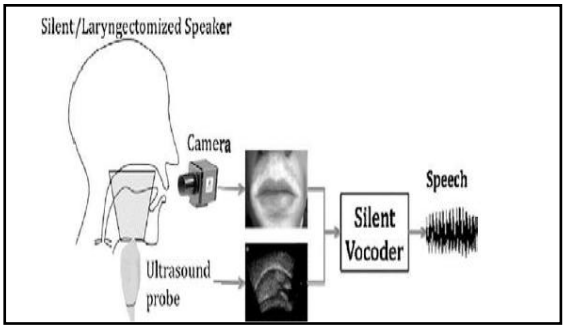


**Fig 3 Schematic of ultrasound silent sound interface[13]**

The speech synthesis from the analysis of articulator motion is usually done by the use of a two- or three-dimensional articulator model of the vocal tract However, the state of the art in high quality speech synthesis uses a segmental approach (or HMM-based methods, as discussed later), in which the speech waveform is obtained by concatenating acoustic speech segments. This technique uses visual observations of articulators to drive a Segmental Vocoder. The terms "visual information" and "visual observations" are taken to refer both to the ultrasound and optical images. This approach integrates a phone recognition stage with a corpus-based synthesis system. As the first (off-line) step, an "audio–visual" dictionary is built from a training set containing units which associate the acoustic and visual realizations of each phone; this dictionary will later be used in the synthesis step. Given a test sequence of visual information only, the vocoder generates the speech waveform in three stages as shown in figure 4:

- An HMM-based decoder predicts a sequence of phonetic targets from the given set of visual features.
- A unit selection algorithm, driven by this prediction, searches in the dictionary the optimal sequence of audio–visual units that best matches the input test data.
- A speech waveform is generated by concatenating the acoustic segments for all selected units.

Since there is no glottal activity, recovering an " acceptable" prosody from "silent data" is an issue, and prosodic transformations of the synthesized speech waveform are needed. These transformations are achieved using "Harmonic plus Noise Model" (HNM)

The approach investigated is based on the construction of a large audio-visual unit dictionary which associates a visual realization with an acoustic one for each diphone. In the training stage, visual feature sequences are modeled for each phonetic class by a context-independent continuous Hidden Markov Model (HMM). In the test stage, the visuo-phonetic decoder "recognizes" a set of phonetic targets in the given sequence of visual features Evaluated on a one-hour continuous speech database, consisting of two speakers (one male, one female, native speakers of American English), this visuo-phonetic decoder is currently able to correctly predict about 60 % ofphonetic target sequences, using video-only speech data. At synthesis time, given a phonetic prediction, a unit selection algorithm searches in the dictionary for the sequence of diphones that best matches the input test data, and a "reasonable" target prosodic pattern is also chosen.The speech waveform is then generated by concatenating the acoustic segments for all selected diphones, and prosodic transformations of the resulting speech signal are carried out using "Harmonic plus Noise Model" (HNM) synthesis techniques. An overview of the segmental approach to silent vocoding is given in figure 4.[13]

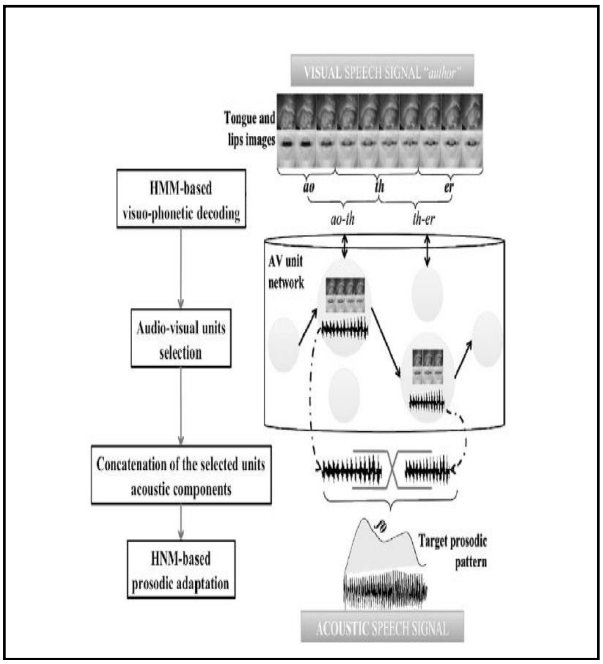


**Figure 4 Overview of the segmental approach for a silent vocoder driven by video-only Speech data.[13]**

1.3.1. DATA ACQUISITION [14]

As shown in figure 5, the hardware component of the system is based on:

- The Terason T3000 ultrasound system which is based on a laptop running Microsoft Windows

XP and provides 640x480 pixels resolution images
- 140° microconvex transducer with 128 elements (8MC4)
- An industrial USB color camera able to provide 60 fps with a 640x480 pixels resolution (USB 2.0, WDM compliant)
- An external microphone connected to the built-in soundcard of the *T3000*

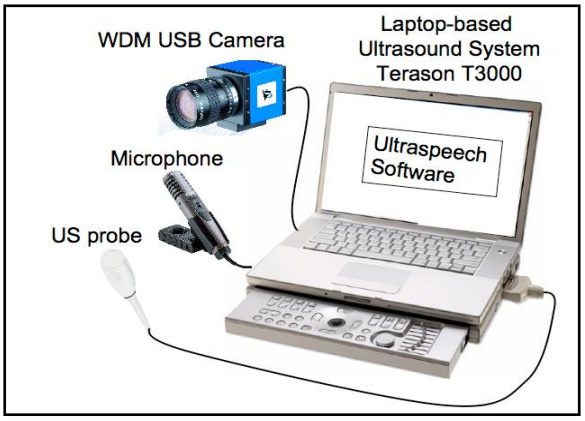


**Figure 3 shows a typical ultrasound SSI system**

In the context of a silent speech interface based on tongue and lip imaging, the desired acquisition system should be able to record synchronously ultrasound data and video data at their respective maximum frame rate together with the acoustic speech signal. In order to have a compact, transportable, and easy-to-use system, a PC-based hardware architecture coupled with a single control program has been adopted. In the described system, data streams are recorded, processed and stored digitally on a single PC using our stand-alone software Ultraspeech.

## III. COMPARISON OF TECHNOLOGIES

The research article titled "Silent Sound Interfaces"[15] contrasts these two technologies based on the following parameters:
- Works in silence – Can the device be operated silently?
- Works in noise – Is the operation of the device affected by background noise?
- Works for laryngectomy – Can the device be used by post-laryngectomy patients? It may be useful for other pathologies as well, but laryngectomy is used as a baseline.
- Non-invasive –Can the device be used in a natural fashion, without uncomfortable or unsightly wires, electrodes, etc.?
- Ready for market – Is the device close to being marketed commercially? This axis also takes into the account in a natural way the current technological advancement of the technique, responding, in essence, to the question, ‘‘How well is this technology working as of today?”.
- Low cost – Can the final product be low cost? The answer will depend, among other factors, on whether any ‘‘exotic” technologies or procedures are required to make the device function.

The article discusses the results(Scores out of 5) as follows:

### 1.4. ELECTROMYOGRAPHY

**Works in silence**: 5 – Silent articulation is possible.
**Works in noise**: 5 – Background noise does not affect the operation.
**Works for laryngectomy**: 5 – No glottal activity is required.
**Non-invasive**: 4 – A facemask-like implementation should eventually be possible, thus eliminating unsightly glued electrodes.
**Ready for market**: 3 – EMG sensors and their associated electronics are already widely available.
**Low cost**: 4 – The sensors and the data processing system are relatively manageable.

1.5. ULTRASOUND SSI

**Works in silence**: 5 – Silent articulation is possible.
**Works in noise**: 5 – Background noise does not affect the operation.
**Works for laryngectomy**: 5 – No glottal activity is required.
**Non-invasive**: 4 – Miniaturisation of ultrasound and camera and gel-free coupling should eventually lead to a relatively portable and unobtrusive device.
**Ready for market**: 3 – Recognition results suggest a useful, limited vocabulary device should not be far off, but instrumental developments are still necessary.
**Low cost**: 3 – Although costs much below those of medical ultrasound devices should eventually be possible, ultrasound remains a non-trivial technology.

## IV. FUTURE PROSPECTS

As already discussed, Electromyography interprets the speech merely by analysing the signals generated by monitoring the muscle activities in the vocal tract. Nanobots can be injected into the blood stream. These bots, attached with RNA strands, adhere to skeletal muscles of vocal tract and transmit information about vocal activity. Instead of employing chords to measure electric signals, these signals can be monitored through special sensors attached to mobile or portable devices and help transmit information wirelessly.

TERC band is interfaced with cellular phone: Transmission of the acoustic signal is

managed via an application. The TERC band transmits the vibration signal developed due to glottis movements.

**Computer brain interface**: Brain signals are sensed by sensors and then they are processed and compared with the patterns stored in database and accordingly the speech is converted.
Image Processing Techniques like Image Segmentation and Histogram Equalisation can be used to compare the images of tongue and lips obtained with the database. There has been a revolution as far as techniques for image comparison are considered.

## V. CONCLUSION

As we can see from the survey results mentioned in Section 3, both the technologies can work in noisy environments to produce desired results. These technologies are non invasive and safe on medicinal grounds. With the advent of nanotechnology, nanobots can be assigned the task to retrieve the signals of vocal activity, giving away with the chords required to monitor. Also the entire system can be turned wireless by sending the monitored signals wirelessly for analysis. Speech Synthesis techniques have come a long way since LPC method. Also, Mobile manufacturing companies can have a large chunk of profits if they embed these technologies in the cell phones. These advantages are indicative of the fact that Silent Sound Technology has it all for it to be embedded in our daily lives.

## REFERENCES

[1] Shehjar Safaya, Kameshwar Sharma, Silent Sound Technology- An End to Noisy Communication, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 9, November 2013

[2] Karishma Gupta, Mohd. Salman Khan, SILENT SOUND TECHNOLOGY , VSRD International Journal of Computer Science &Information Technology, Vol. IV Issue I January 2014

[3] Petajan, E.D., 1984. *Automatic lipreading to enhance speech recognition*. IEEE Communications Society Global Telecommunications Conf., Atlanta, USA

[4] Nakamura, H., 1988. *Method of recognizing speech using a lip image*. United States Patent 4769845, September 06

[5] Sugie, N., Tsunoda, K., 1985. *A speech prosthesis employing a speech synthesizer-vowel discrimination from perioral muscle activities and vowel production*. IEEE Trans. Biomed. Eng. BME-32 (7), 485–490

[6] Hasegawa, T., Ohtani, K., 1992. Oral *image to voice converter, image input microphone*. Proc. IEEE ICCS/ISITA 1992 Singapore, Vol. 20, No. 1, pp. 617–620.

[7] Ng, L., Burnett, G., Holzrichter, J., Gable, T., 2000. *Denoising of human speech using combined acoustic and EM sensor signal processing*. In: Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Istanbul, Turkey, 5–9 June 2000, Vol. 1, pp. 229–232.

[8] Nakajima, Y., 2005. *Development and evaluation of soft silicone NAM microphone*. Technical Report IEICE, SP2005-7, pp. 7–12 (in Japanese)

[9] Kamen, Gary. Electromyographic Kinesiology.In Robertson, DGE *et al*. Research Methods in Biomechanics. Champaign, IL: Human Kinetics Publ., 2004.

[10] Huei-Ming Chai, *Measurements of Muscle Strength*, www.pt.ntu.edu.tw/hmchai/Biomechanics/BMmeasure/MuscleStrengthMeasure.htm

[11] HTLab, Physiological Lab, http://htlab.psy.unipd.it/index.php?page=physiological-lab

[12] Arthur R. Toth, Michael Wand, Tanja Schultz, *Synthesizing Speech from Electromyography using Voice Transformation Techniques*, Interspeech Brighton, 2009, http://www.cs.cmu.edu/~atoth/papers/IS090521.pdf

[13] B. Denby, T. Schultz, K. Honda , T. Hueber , J.M. Gilbert, J.S. Brumberg, Silent Speech Interfaces, Speech Communications, 2009

[14] T. Hueber, Acquisition of Ultrasound, Video and Acoustic Speech Data for a Silent-Speech Interface Application, IEEEXplore

[15] B. Denby, Silent Speech Interfaces, Speech Communications,www.elsevier.com/locate/specom'