Current Techniques Of Silent Speech

Clavijo Juan, Olga Ramos, Dario Amaya

Military University New Granada

Received 12 July 2016; Accepted 28 August 2016; Published 1 September 2016

ABSTRACT

Silent speech is a technique based on noticeable information in the vocal tract, in the absence of sonorous signals. This article is a review of previous investigations about the silent speech area, highlighting the four most published acquisition techniques: Electromyography (EMG), not audible murmur (NAM), electromagnetic fields and image diagnostic. This review is based on 51 references including some indexed journal articles and current technological literature, evaluating its applications, techniques and mathematic methods. As a result, this review concludes the strengths and weaknesses of each analyzed techniques, through a comparison chart.

KEYWORDS: Electromyography (EMG), not audible murmur (NAM), electromagnetic fields and image diagnostic.

INTRODUCTION

For humans, vocal speech represents the most common sharing information system between them, however some circumstances make sonorous communication impossible, for example when some abnormality occurs in the vocal tract system, in high noise pollution situations or underwater environments among others. In those circumstances sub vocal speech arises as an alternative way of communication. The silent speech techniques are an important part of the most ambitious technological researches of the last two decades [1][2][3][4]. This techniques use electromyography, non-audible murmur, images, cameras, ultrasound equipment, motion readings based on electromagnetic fields variations, whiling to identify letters, syllables, words and phrases [4][5]. Silent speech systems are also an alternative to Man –Machine communication [6][7][8][9][10]. Nowadays this communication is possible with human interface devices like keyboards and microphones among others [9][11][12]. Systems of vocal speech recognition arises as the starting point for this investigations [13][14][15]. Thus, a computational algorithm can recognize the speech and make a written representation of it, this technique uses the vocal tract to create the right language pronunciation. The sub vocal speech aims the same purpose but silently. Trunking style communication, is used in high confidentiality and stealth military operations [16], or in mobile phones hands-free systems, which are also feasible sub-vocal speech applications. In some other cases it can be used to give instructions to computationally controlled systems, or even applications where robot activities are controlled and executed through commands. The silent speech is not audible, which implies that people who use it, don’t use the most of their vocal tract [17]. Non audible effect can be formed just by thinking about the letters, language or the words, this action produces nerve stimulation over the vocal tract muscles [1][2]. The sub vocal meaning is also associated to murmuring which is the words pronunciation avoiding to open the mouth and limiting the joints of the oral cavity. The small stimulations can be detected as electromyography readings [4][5]. Low level readings are also measurable when the vocal tract system is partially used, creating the concept of non-audible murmur [18], which consist in the articulation and use of the vocal tract system without pronouncing the words loudly or not even making sounds [4][17][19][20]. A parallel alternative is by scanning, acquiring images of the vocal tract system by means of video cameras or...
ultrasounds [4][21][22]. This techniques make the images analysis to determinate user’s voice patterns. It is also possible to track and identify the vocal tract movement by means of oscillation electromagnetic field sensible transducers. This arrangement of sensors is set with field permanent magnets in specific positions over the vocal tract or the face [4][23]. In first instance this review shows how the human voice is produced, contextualizing some vocal tract system concepts such as voice speech, silent speech and sub vocal speech. Also the relation between the vocal tract system and brain nervous interaction. Then, the greater impact techniques for silent speech identification are illustrated, highlighting the image analysis, motion detection with magnetic sensors, electromyography readings and no audible murmur.

This Article concludes, with some conclusions and a comparative analysis of the presented techniques, made by means of a set of five characteristics that relates them.

Methodology:
The Military University New Granada, in exercise of its research politics, proposes and develop many different scientific investigation projects. One of them is the Recognition Interface of the Spanish alphabet through non audible murmur (NAM). That investigation is where this review article comes from, this revision is obtained by doing an exhaustive search of bibliographic material, in current databases and indexed journals.

2.1 Human Voice Formation:
The human being creates the phonemes by a physiological system called vocal tract [24]. The phonemes are created by the mutual interaction of several system parts. First of all an airflow is produced by the diaphragm and the lungs this airstream or airflow flows up through the trachea and collides with the vocal cords. The cords are membranes with the capability to open or close the gap between them. Simultaneously, the vocal cords can also change them tension like the strings of a guitar when tuned with pegs. This motion game of opening and closing produces vibrations, creating all the different frequency ranges that the human voice can have. Subsequently, words and letters articulation is achieved by movements of the jaw, tongue and lips [13]. The average population human range frequencies is 300 to 3K Hz. The vocal tract system movements are governed by many muscles contraction and relaxation. However this muscles doesn’t work without the respective brain nervous stimulus. The human brain has two areas where it processes and administrate the human speech. This zones are named Broca and Wernicke. [25][26][27]. According to scientific research made by Paul Pierre Broca, French doctor and Karl Wernicke a German neurologist and psychiatrist, the Broca’s area, is the portion of cerebral cortex in charge of vocal’s tract movement control [27]. Wernicke’s area allows the capacity of articulating words and its formation in phrases and sentences, which means that they have a mutual relation for fluid speech pronunciation. Current researches, evidence brain activity in other areas among the Broca’s and Wernicke’s. One related study in [28], shows new zones associated with the language. The physiological concept mentioned above, is articulated by means of brain nervous connections between the brain and the vocal tract system muscles, this allows the the muscular movements to create the phonemes.

Speech Identification By Image Analysis:
The diagnostic generation imaging by echography, are made from ultrasonic lectures. Based on this principle, there are simpler variations which don’t require physical contact with the user [29][30][31]. Images has been really helpful in phonemes identification. Its application is made in two different ways, the first is the video acquisition over the vocal tract system, specifically the mouth [4] and the second is based on ultrasound scanning ultrasonic [32]. In the first case, the mouth movements are recorded by a video camera and processed to identify letters or phonemes [33]. This is based on contour detection in the image, to define the mouth size. Investigations on this technique have identify the pronunciation of letters such as vowels. To do so, they use lips contour dimensional conditions like, height, width and area [4][33]. The second way uses and ultrasonic scanner, to illustrate images where organs and tissues can be observed. Unlike the case above, this technique identifies the characteristics through visual analysis, showing the area, curvature and shape, that the tongue assumes pronounce a phoneme [4][21][22]. The implementation of this technique is shown in figure 1.
Speech Identification By Magnetic Sensors:

The magnetic field is done using Hall Effect sensors [34]. This sensor shows the present magnetic fields by voltage variations through its output. The field readings can be made in three axes, it depends on the device. When a Hall Effect sensor is exposed to moving permanent magnets, its readings can be characterized. And that’s how is possible to track the mouth muscle movements by installing permanent magnets [4][23][35]. Using a set of this sensors is a way to detect the magnetic field in different directions, a precise example of this application are the compasses and digital compasses, which can detect weak magnetic field of the earth, identifying the polar orientation [4]. Research [23], implemented six Honeywell HMC 1022 sensors, to measure the magnetic fields, generated by constant magnets, the distribution of sensors and magnets is shown in figure 2.

Fig. 1: Ultrasonic Scanner [22].

Fig. 2: Hall Effect Sensors [23].

In [23], 12 bit resolution and 4K Hz sample frequency is used, the collected data was acquired by an application programmed in MATLAB, capable of identify some phonemes such as “CAT” and “DOG”. Statistical characteristics of the acquired signals were tabulated and classified for identifying patterns.

2.4 Speech Identification By Means Of Electromyography Signals:

Electromyography signals, are electrochemical brain stimulations located for muscles [4]. This signals allows the muscle contraction. Due to the signal electricity, it is possible to acquire a sample using biomedical instrumentation [36], using the right electrodes. The electromyography readings can be invasive or noninvasive, it depends on the chosen muscle complexity [1][2][37]. Taking advantage of the electromyography readings capacity, studies and developments has been done, allowing the identification of vocal tract muscles movements. The variety of the acquired data differs from one research to another, using different quantities of noninvasive electrodes [1][2][5][37][38][39][40]. The ability of read slight muscle electric signals gives the chance of knowing when a the subject is thinking about the phoneme and is ready to say it [1][2][4]. When this happens, the vocal tract muscles, receive electromyography to initiate the execution of movements, but it doesn’t necessarily end with the speaking action. The investigation [39], allows the silent speech recognition, by means of eight electrodes, located over the face, according to the speech muscles positions. The acquired electromyography images are processed by a finite state machine, the machine detects the speech start and then it runs the identification analysis based on Markov models techniques. The ten digit numbers were used as test patterns [39]. Aiming to recognize silent speech, in [38] working comparatives were proved between vocal and silent speech readings. This analysis was accomplished using the phrase “HELLO WORLD” 78 samples, over 92 sessions were used. A total test time of 395 and 361mins, for vocal and silent speech para respectively. The
conclusion is the phoneme modeling in function of pronunciation and the related muscles activation, and the analysis of the alveolar, glottal, explosive, fricative, vocal and frontal. The details of this concepts can be seen in [41]. A combination of the characteristics mentioned above can be present in one phoneme. This information is analyzed statistically to predict or identify the pronounced phoneme [38]. One of the most representative works about sub vocal speech it’s been developed by NASA [1][2][4], the investigation result is the sub vocal speech identification through electromyography readings, using two noninvasive electrodes. And the recognition strategy is the Wavelets threes transform. This investigation has been applied to technologies of man-machine communication, for robots and words recognitions. [1][2]. Figure 3, presents a demonstration of the NASA system.

Fig. 3: Sub-vocal speech NASA system [2].

Speech identification through non audible murmur:
The non-audible murmur, is presented when the acoustic power is low or null. The non-audible murmur is completely mute when the muscles are articulated, but there’s not airflow from the lungs. The murmur could be slightly audible, when there’s a small airflow that can be heard as noise [42]. The non-audible murmur is acquired by specialized microphones, and located over the skin surface, the strategic location is close to the ears behind the pavilion or over the sternocleidomastoid muscle [19][42][43]. The NAM microphones works as the medical stethoscopes [17][18][19][44][45], by means of a bell and diaphragm incorporating an electret microphone inside. The fusion between the bell and the electret microphone made by a silicone surface. Weak muscular movements generated by the vocal tract, are acquired by the NAM microphone, Figure 4 illustrates the typical microphone location.

Fig. 4: NAM Microphone [43].

Investigation [18], uses the NAM microphone as a silent speech sensor, showing a method also for its construction, the position over the skin surface is also studied, choosing the best location to optimize the characteristic of the acquired signal [18], a recognition alternative is also proposed by Markov hidden models [46][4] the advantage of this device in conditions of ambient noise such as tv noise. The success of Markov models depends on the election of the right probability tree and its implementation [46][18]. A set of Markov model rules for Spanish language can be consulted in [47]. In [48], the improvement of the recognition Markov models is proposed by changing the distance between the model points [19], set the importance of silent speech, associated to the mobile phone communication systems, highlighting the its restrictions in public places and the application of NAMs as hands-free systems. French alphabet are also analyze, highlighting the importance of speech fundamental frequency detection. This condition improves Markov models identifications and its Gaussian distributions. To determinate the fundamental frequency an artificial neuronal network is applied. The research test was executed with the /a/, /e/, /i/, /o/, /u/ and the consonants, /p/, /pj/, /b/, /bj/, /m/, /mj/ with bilabial
pronunciation. /d/, /t/, /s/, /ts/, /z/, /j/, /n/, alveolar pronunciation /k/, /kj/, /g/, /gj/, palatal pronunciation, among others.

The investigation [17], proposes a speech phonemes recognition system, which are acquired through a NAM microphone. This development analyses the microphone data by means of the hidden Markov models and it also lean on photograms that represents speech spectrograms. They also presents Japanese phonemes /a/, /e/, /i/, /o/, /u/, in a tridimensional space, looking at the main frequency components in the NAM signal, and comparing them with normal audio acquisition. In [49], the way some normal body noises interfere with the quality of the acquired speech data by the NAM microphones. As an improvement method the investigations [49] and [50] proposes the usage of an algorithm of maximum likelihood estimation (MLE). The use Gaussian Markov models, determine the speaker phone fundamental frequency as in [19] and also the frequency analysis of the acquired signals of the NAM microphone. The article related in [51], shows improvement methods for NAM microphone signals. To reduce noise conditions, a mathematical method is proposed to compensate the signals such as spectral media difference (CMS), constrained maximum likelihood linear regression (CMLLR) and maximum with linear regression restrictions (CSMAPLR). Between results and comparisons in investigation [51], it is noticeable that NAM microphone location offers some levels of distortion in acoustic signals. Figure 5 shows the comparatives results between the used techniques, compensation and performance.

![Fig. 5: Techniques comparative [51].](image)

In [51], the implementation of a compensation method and its advantages to improve the speech conditions recognition with NAM systems. Due to the results the CSMAPLR is considered the best method. The investigation [42], proposes speech identifications using normal transform (SAT). Coefficients ant characteristics of SAT system, are achieved with the implementation of a target function and also an optimization criteria. The results of [42] investigation, evidences the possibility of improvement of NAM identification system in 2% thanks to SAT. In [42] a comparative histogram between the basic SAT technique and the experimentation obtained results is presented. In this investigation the performance factor and words improvement (WACC) is presented perceptually. The investigation proposed in [20], presents a technique to reduce blind noise that is produces by the body and its natural movements and process like the breathing and heartbeat. In addition, there are other external noises such as limb movement and friction of clothing, investigation [20], proposes the implementation two NAM microphones, creating the stereo effect in sound to reduce them.

**Comparative Analysis:**

The current techniques for silent speech acquisition, have several forms in function of the strategy and the method over the vocal tract system. The strategies mentioned above, images, magnetic sensors, electromyography and non-audible murmur, has become the most popular in the last decades, the last two, capture grater interest than the others. Thus to their silent speech recognition results. The characteristics implemented in EMG and NAM, ease its usage and they are also the guideline to global investigations. The mathematical strategies implemented for this two techniques are many, like, FFT[20], SAT [42], HMM [46], MLE [49], CMS[51], CSMAPLR [51]. It should be noted that the target of all the investigations reviewed in this article is the same, silent speech phoneme identification as precisely as possible. The particular objectives in each aspect differs in aspects like: vowel phonemes, words and medical conditions identification. This characteristics allows this developments to be useful in purposes like: robotics, electronic control, physical and electronic communications. The EMG and NAM implemented techniques, are the best way to develop silent speech recognitions portable systems. Low hardware requirement over the vocal tract, proliferates this investigations inside research groups and centers. Table 1, relates a comparative analysis, quantifying the
aspects on a scale of 1 to 5. The compared aspects were selected in function of the generalities that fall within the four techniques discussed throughout this article

<table>
<thead>
<tr>
<th>Criterion of analysis</th>
<th>CV</th>
<th>NAM</th>
<th>EMG</th>
<th>MAG</th>
<th>IMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>User convenience</td>
<td>+</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Easy Installation</td>
<td>+</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>System portability</td>
<td>+</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Hardware resources or equipment</td>
<td>-</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Processing requirements</td>
<td>-</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

From the analyzed aspects in table 1 some can be highlighted: User convenience is about the user physical comfort when using the recognition system. In this feature a five, represents a high comfort system and a rating of one, evidence a low user comfort. Easy Installation, is a comparison parameter over the easiness in the installation process. This means that if the rating assigned is five, the system is easy to install and if it is one, then it is too difficult. System portability, this feature evaluation quantifies how portable is the implemented system. A five, denotes that the system is portable, however a rating of one defines that the system is difficult to transport. Hardware resources or equipment, represent the number of physical computers, electronic cards and infrastructure required to make the system operational. A rating of five indicates that the amount of these resources is high, while a score of one indicates that this resource is scarce. Processing requirements, this approach presents an indicative that defines the requirements of computational load, like processing speed, RAM and ROM, required to make the recognition system work properly in real time. Five rating, defines the required computational load is high, quantification of one defines the required load is low. Table 1, compares the five chosen criteria, as a way to evaluate the exposed techniques limits, pros and cons. This article also presents a weighted average that quantifies the performance of each technique in function of the criteria. The second column CV, illustrates a valuation criterion that can be positive or negative. This valuation criterion, indicates how they affect the ratings on the average valuation. Finally, the comparative analysis presents in its results, the scores are generated based on the evaluation criteria, showing that the best techniques are the NAM and EMG respectively.

Conclusions:
The analyzed techniques in this article, evidences the pertinence of the NAM and EMG systems usage. Thus their portability and the small amount of hardware required. Regardless the applied technique, is important to use high computational load mathematical analysis systems in software and hardware due to the required heavy signal treatment. It would be accurate to carry on the research about NAM microphones, and its experiments with stereo systems, because this technique filter blind noises of the human body, optimizing the recognition process. Speech identification systems adaptation is closely linked to the characteristics of each speaker, such as fundamental voice frequency and pronunciation speed.

REFERENCIAS