Applied Phon Curve Algorithm for Improved Voice Recognition and Authentication

B.L. Tait

University of Johannesburg, South Africa
bobby@cs.uj.ac.za

Abstract. The ability of a robot, computer or any man made system to understand exactly what a human, and who the human is that said it, is the focus of many research projects. IBM Via voice [1], and efforts in the Microsoft XP operating system, endeavoured on understanding what a person said. We cannot argue the fact that it would be fantastic if a PC can listen, interpret and understand what a human commands. However, this type of effortless, exact voice commanding is still only a feature experienced in futuristic stories like Star Trek. This paper considers a novel approach in improving the current voice recognition and authentication efforts in existing software systems. It does not replace or make any current efforts absolute. In this paper the way that sound essentially works is discussed; Research by Fletcher-Munson [2], [3] on equal loudness is integrated into a new voice recognition proposal, and implemented as a middle tier software algorithm. Considering the suggestions and findings of this paper, will serve as a stepping stone towards allowing man made systems to interact with humans, using voice commands. The application of this algorithm improves the false acceptance rate and false rejection rate of tested voice authentication systems.

Keywords: Voice recognition, Biometrics, Security, Phon Curve, Characteristics of Sound, Fletcher-Munson, Voice authentication.

1 Background

2001: A Space Odyssey is an old movie filmed in 1968 [4]. The central character of this movie is a central command computer known as HAL. The impressive feature of HAL, among many others, is the ability of HAL to seamlessly communicate with the crew on the space ship, using spoken word.

Interacting with a human in this way, today, nearly forty five years later, is still science fiction. Even if you really make a lot of effort to "educate" the software, the system will still often make the odd misinterpretation of what was actually said.

Sound technology, on what all voice recognition and voice authentication approaches are based, is in no way a young technology. The ability to transduce sound into electricity is based on magnetic induction, discovered my Michael Faraday in 1821 [5].

To store sound, analogue methods were first developed, which, among other approaches, physically etch the vibrations from the transducer, into a medium like tin foil, developed by Thomas Edison [6].

H. Jahankhani et al. (Eds.): ICGS3/e-Democracy 2011, LNICST 99, pp. 23–30, 2012.

Today, sound is digitized using a pulse code modulation (PCM) algorithm, to store or manage sound signals. An IPod is a prime example of a sound device relying on digitized sound.

2 Introduction

This section will briefly consider the building blocks of sound, which is used to identify sounds produced by humans.

2.1 Transduction

A computer system cannot directly interpret a voice commands, it must be digitally presented, for the computer system to work with the sound from this voice.

Thus sound must be changed from sound energy into electrical energy; this is done using a microphone. The first step is to present the sound as electricity (which is analogue in nature). The next step is to convert the analogue electrical signal into a digital representation of the sound. Once in digital format, software algorithms can be applied to process the sound, and analyze the sound. The process of changing sound into electricity is accomplished by two main methods [7], [8]:

Dynamic transduction relies on magnetic induction to generate an analogue electrical representation of a particular sound. A picture of a dynamic microphone is illustrated in figure 1. Essentially sound waves move the diaphragm which is attached to a voice coil, seated on a magnet. The sound waves cause the diaphragm to vibrate, vibrating the voice coil, which, because of the magnet, induces an electric current on the voice coil. This current is then sent to a circuit for further amplification (using a microphone pre-amp).

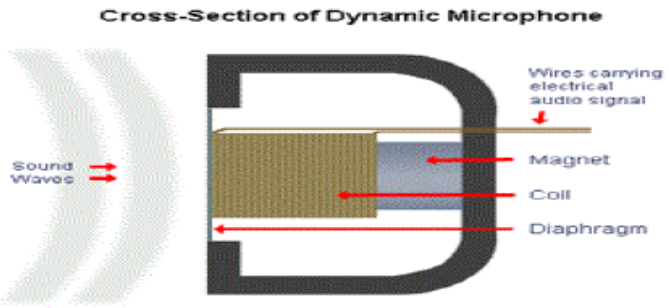

Fig. 1. Cross-Section of dynamic Microphone

Condenser transduction relies of static electricity to transduce sound into an analogue electrical representation of the sound. A picture illustrating the working of a condenser microphone is illustrated in figure 2.

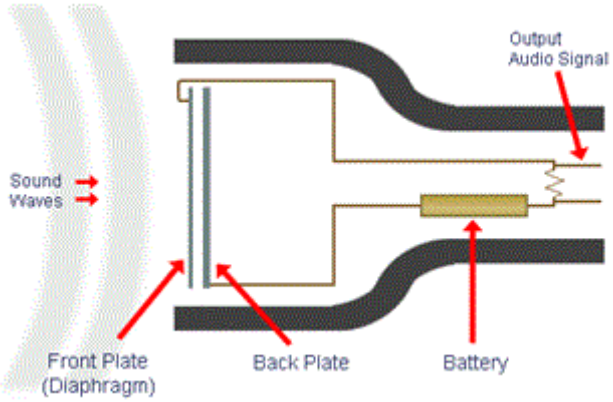

Fig. 2. Condenser transduction

In short the system relies on two plates to be charged with a static electric load. Let's assume that the two plates are charged with a negative static load. If the sound wave are to exert sound pressure on the top, thin plate (in high quality microphones, made of thin gold foil), the negatively loaded plate would move closer to the negatively charged bottom plate, causing electrons to flow through the circuit, thus once again creating an analogues flow of current through the circuit. The transduction is a very important part of the chain of sound production. If the microphone does not effectively convert all the sound signals to electricity, the transduced sound will not be a true representation of the sound that actually occurred. The biometric decision algorithm relies on very specific tolerances, to make the correct decision. If a low quality microphone is used, the false rejection rate is often negatively influenced. If voice recognition is the aim of the system developed, Condenser transduction should be favoured over dynamic transduction. The reasons for this, falls beyond the scope of this paper.

3 Sound Characteristics

Sound characteristics are the fundamental tools used in voice recognition; sound is composed of 3 characteristics [9] – Pitch, Dynamic range, and Timbre. Whenever a human speaks, these 3 characteristics will always be present, in fact for any sound that exists, these 3 characteristics will be present.

3.1 Pitch

As sound is cyclic vibrations, which causes a medium to vibrate, the cycle of each vibration is measured in Hertz (1Hz is one full cycle in one second). A young human ear is sensitive to hearing sound vibrations ranging from 20Hz to 20 KHz. The fundamental sounds that humans produce when speaking is roughly between 120Hz and 7 Khz (take note that there will be a lot more sound content produced by a human, but will be elaborated on, later in the paper). The more accurate the

microphone can transduce the sound vibrations emanating from a human, the better the electrical representation will be of the actual sound.

3.2 Dynamic Range

The second characteristic is dynamic range or loudness. A human year can hear a wide dynamic range. The difference between the softest audible sound and the loudest is exponential in nature. If we assume that the dynamic range of two fingers brushing against each other is for e.g. a measure value of 23, then the sound of a rocket being launched should be around 4,700,000,000. In order to address this, Alexander bell devised the decibel [11]. Decibels are designed for talking about numbers of greatly different magnitude, thus huge deviation found in all possible values, such as 23 vs. 4,700,000,000,000. With such vast a difference (deviation) between the numbers, the most difficult problem is getting the number of zeros right. We could use scientific notation, but a comparison between 2.3 X 10 (23) and 4.7 X 10 to the 12th (4,700,000,000,000) is still awkward. In sound the decibel must be presented as a unit of specific measurement, for example decibel SPL (sound pressure level) or decibel Watt, depending on what aspect of the dynamic range was measured.

The decibel of interest in the paper is decibel SPL, which can be calculated using the following formula: 20 Log(Measurement A / Measurement B).

Where the measurement A will be the loudness of an initial sound pressure, and measurement B is the second, altered sound pressure measurement. According to this calculation, the Db SPL of finger brushing against each other will be close to 0 Db SPL (Zero Decibel SPL) [10], and the sound of a launching rocket will be around 140 Db SPL [10]. Normal conversation should be around 60Db SPL [10]. Take note that every 6 Db SPL will sound like a doubling of the SPL on the point of transduction (however, if Db Watt is to be considered, every 3db Watt, will be a doubling of for example current consumed or produced).

3.3 Timbre

The last sound characteristic is known as timbre. Timbre is also referred to as harmonics or overtones. Timbre is determined by its spectrum, which is a specific mix of keynote, overtones, noise, tune behavior, envelope (attack, sustain, decay), as well as the temporal change of the spectrum and the amplitude [11]. As this is the characteristic in sound, mostly responsible for uniqueness, this aspect of sound should enjoy the greatest attention, if voice authentication is to be considered.

Sound is only composed of vibrations (pitch) and loudness (dynamic range). Timbre is also only vibrations, emanating from the source. If we take for example two males, and we ask them to sing a A4 note (440Hz) at the same loudness (let's say 60 db spl), we will agree that the vibrations they both create is around 440Hz, and they both sing the same loudness, but they sound clearly different. Timbre is the reason for this difference. Timbre describes those characteristics of sound which allow the ear to distinguish sounds which have the same pitch and loudness. If any object vibrates in a medium, the object will have a fundamental vibration, but this is not the only vibration which will occur. A human's chest cavity, nose cavity and many other body

parts will vibrate and resonate because of the fundamental vibration. The harmonic series is important in objects which produce sounds. The natural frequencies of the string mentioned above form a harmonic series.

A frequency is harmonic if it is an integer multiple of the fundamental frequency. The fundamental is the first harmonic (although it's generally referred to as the fundamental). The second harmonic is two times the frequency of the fundamental; the third harmonics is three times the fundamental, and so on. So with a fundamental of 100 Hz, the second harmonic is 200 Hz, the third is 300 Hz, the fourth is 400 Hz, or if the fundamental frequency is 25Hz, the second harmonic is 50 Hz, the third is 75 Hz, the fourth is 100 Hz etc. Due to harmonics, the frequency range of the human voice can run from 42Hz right up to 30Khz [12]. Often the digitizing process (PCM) will only convert 20hz to 20KHz to digital format.

4 Equal Loudness

Fletcher-Munson [2], [3] conducted research on the way that humans actually hear. They determined that humans hear frequencies differently in relation to other frequencies based on the dynamic range of the frequency. During their research, they tested many humans. Equal-loudness contours were first measured by Fletcher and

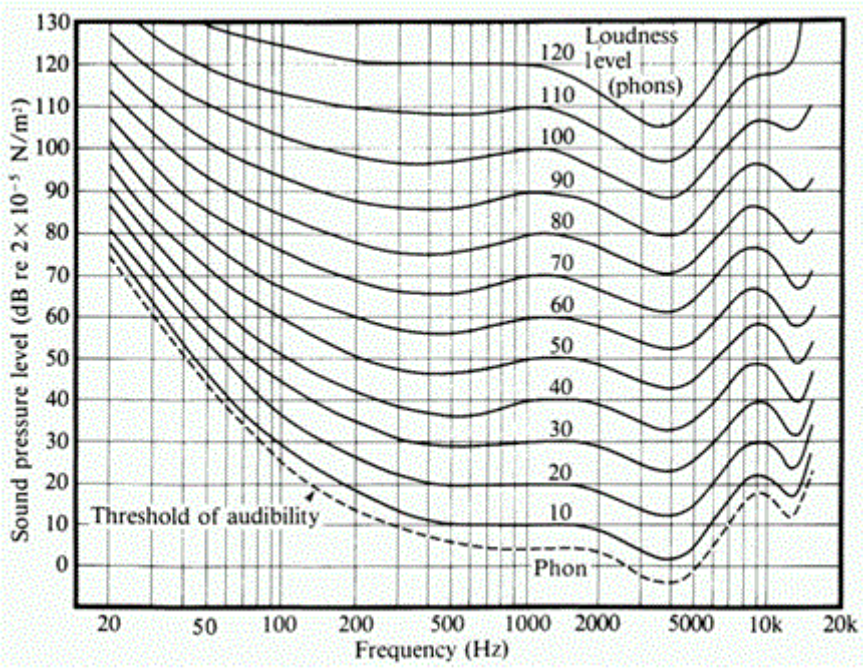

Fig. 3. Phon Curves

Munson using headphones. In their study, listeners were presented with pure tones atvarious frequencies and over 10 dB increments in stimulus intensity. For each frequency and intensity, the listener was also presented with a reference tone at 1000 Hz. The reference tone was adjusted until it was perceived to be of the same loudness as the test tone. Loudness, being a psychological quantity, is difficult to measure, so Fletcher and Munson averaged their results over many test subjects to derive reasonable averages. From this research they produced the Fletcher-Munson curve, illustrated in figure 3. The Fletcher-Munson curve is a measure of sound pressure frequency for which a listener perceives a constant loudness. The unit of measurement for loudness levels is the phon, and by definition two sine waves that have equal phons are equally loud.

Our ears do not perceive all sounds equally at the various frequencies or sound intensities. The sound levels for a particular sound as defined by the level at 1000 Hz will find the same for any given frequency along the curve. This indicates that our ears are less sensitive to low frequency sounds than mid to high frequencies.

The next section of the paper will consider the proposed application of the Fletcher-Munson curves.

5 Applied Phon Curve for Voice Recognition and Authentication

It is clear from the aforementioned sections that there are many factors to be considered when working with sound. During this research, presented in this section, the information presented by the Fletcher-Munson curves are used to improve the ability of software to recognize the person speaking, and to determine what the person said.

5.1 Factors to Note Based on the Fletcher-Munson Curves

Frequency spectrum sensitivity: If sound is transduced by a microphone to electricity, a decent microphone will not perceive any difference between for e.g. a 800hz Sound and a 4.5KHz sound. However, a human will clearly hear frequencies between 2KHz and 6Khz much clearer, and better than the other frequencies found in the audible spectrum.If a computer needs to “listen” the way humans interact, the computer must weight the importance of frequencies.

Bass frequency sensitivity: A second observation if the Fletcher-Munson Curves are considered is the fact that a microphone will not perceive any difference between the intensity of bass frequencies (roughly under 500Hz) and other frequencies.

During a test conducted in this research, it can clearly be demonstrated that if a microphone is supplied with let’s say a 40Hz tone, set at 20 db SPL, the human will struggle to hear the test tone, however the microphone will indicate that a 20 db SPL signal is received. On a next test, a test tone of 1 KHz was used, also at 20 db SPL. To the human the 1 KHz tone was clearly audible. In both cases, there was no difference shown by the microphone, as in both cases the microphone register a 20 db SPL signal. However, as indicated by the Fletcher-Munson Curves, if the bass signal

produced is of high intensity, the difference between bass and higher frequencies are not as much as in a the case of bass at lower intensities. If the slope on the bass side (20hz to 500hz) of the 20 Phon curve is considered, and compared with the slope of bass side of the 100 Phon curve, the reader will note that the bass side slope from higher intensities (higher Phon curves) are not as steep as in lower intensities (lower phon curves). The human thus struggles to hear low frequencies with low intensity.

6 Software Algorithm Developed

A software algorithm has been developed to apply the insight provided by the Fletcher-Munson curves.

6.1 Frequency Spectrum Sensitivity

The phon curves are used as a weighting system in the algorithm to adapt the system's ability to "hear" the way the human hears. Due to the fact that humans are unconsciously aware of the frequencies which we hear best, humans tend to accentuate the frequencies during speech. Humans tend to use different frequencies based on different situations. If a lady is in distress, she will use higher pitch frequencies to attract attention. Unconsciously we know that higher frequencies are heard well. This ability is programmed into the algorithm, to allow the system to focus on the frequencies for humans of note.

6.2 Bass Frequency Sensitivity

Secondly the algorithm adapts the loudness based on the phon curve detected when a speech signal is received. Thus is a person speaks softly, the system will expect that the bass content will contain less usable info, compared to a situation when speech signal is received on higher intensities.

7 Conclusions

In order to improve current voice recognition and verification systems, the process of converting sound into digital, should consider the way that humans interact with each other using speech.

Sound equipment provides a very clinical approach, and does not have the ability to listen to humans like humans listen to each other. Research done by Fletcher-Munson, paved the way to understand how humans interact with one another using sound, based on how we hear.

By considering two major factors, frequency spectrum sensitivity and bass frequency sensitivity, as interpreted from the phon curve diagrams, a algorithm was developed. This algorithm adapts the sound signal to ensure that the signal being sent for further processing resembles a closer match to the way humans actually communicate and authenticate each other.

The software is installed as a middle layer, between the digitizing software and he voice recognition software. Though the final statistics are still being evaluated, it was abundantly clear that the software managed to recognize spoken words a lot better, compared to the ability of the software excluding the phon curve adaption.

References

1. IBM Via Voice, http://www-1.ibm.com/software/pervasive/ embedded_viavoice/
2. Vitz, P.C.: Preference for tones as a function of frequency (hertz) and intensity (decibels). Attention, Perception, & Psychophysics 11(1), 84–88 (1972)
3. Fletcher-Munson Curves, http://hyperphysics.phy-astr.gsu.edu/hbase/sound/eqloud.html
4. 2001: A Space Odyssey (1968), http://www.imdb.com/title/tt0062622/
5. Faraday, M.: Science World, http://scienceworld.wolfram.com/biography/Faraday.html
6. The Inventions of Thomas Edison, History of the phonograph, http://inventors.about.com/library/inventors/bledison.htm#phonograph
7. Davis, G., Jones, R.: The Sound Reinforcement Handbook, Yamaha, 2nd edn. (January 1, 1988)
8. Earle, J.: The Microphone Book. From mono to stereo to surround - a guide to microphone design and application, 2nd edn. Focal Press (November 24, 2004)
9. Everest, F.A., Pohlmann, K.: Master Handbook of Acoustics, 5th edn. McGraw-Hill/TAB Electronics (June 22, 2009)
10. Speaks, C.E.: Introduction To Sound: Acoustics for the Hearing and Speech Sciences, Singular, 3rd edn. (March 1, 1999) ISBN-10: 9781565939790
11. Yost, W.A.: Fundamentals of Hearing, An Introduction 5th edn. Emerald Group Publishing Limited (October 2, 2006) ISBN-10: 9780123704733
12. Boulanger, R.: The Csound Book: Perspectives in Software Synthesis, Sound Design, Signal Processing, and Programming. The MIT Press (March 6, 2000)