Machine Learning for Voice Recognition

Saritha Kinkiri
University of Greenwich
Faculty of Engineering and Science,
University of Greenwich,
Chatham Maritime ME4 4TB, UK
S.Kinkiri@gre.ac.uk

Wim J.C Melis
University of Greenwich
Faculty of Engineering and Science,
University of Greenwich,
Chatham Maritime ME4 4TB, UK
Wim.J.C.Melis@gre.ac.uk

Simeon Keates
University of Greenwich
Faculty of Engineering and Science,
University of Greenwich,
Chatham Maritime ME4 4TB, UK
S.Keates@gre.ac.uk

ABSTRACT

Verbal communication is very important to humans, but performing verbal communication with machines still faces particular challenges. Therefore, researchers are trying to find ways to make communication with a machine more similar to communicating with other people, for which two systems have been identified: speech and voice recognition. While speech recognition has aimed to become speaker independent, voice recognition focuses on identifying the speaker, by looking at the tone of the voice, which is affected by the physical characteristics of that person. This requires one to identify these unique tonal features, to then train a system with this data. Being able to perform this identification well, would also bring benefit to speech recognition by allowing the system to adjust to the characteristics of that speaker and how he/she produces their sounds.

Keywords

Machine Learning, Communication, Voice Recognition, Speech Recognition, Security and Biometric Authentication

1. INTRODUCTION

Recently, more and more applications are being developed to improve interaction between humans and machines, supporting a more "natural" verbal interaction between both. Humans communicate with others using gestures, writing text, drawings, facial expressions, and body and sign language [1]. However, verbal is a very important medium of human communication. With regards to interaction with machines, this has however been fairly limited, because, till date, humans adapted to the machine using their language, whereas it would be more beneficial if human's can interact with them in a more natural way.

Research in the area of more natural interaction between machines and humans has focused on developing two systems, which are speech and voice recognition, that allow a machine to either understand and recognize words [2] or identify who is talking, respectively. Creating the artificially intelligent applications [3] that offer speech and voice recognition functionalities would be able to aid humans in many ways to help towards making human life better.

Currently, there is a rising trend to make information, including personal information, such as health related info, available through the internet [4], which increases the risk of fraud and/or identify theft, making security become one of the biggest problems of this technology [5]. Till date, solutions include the use of a password or magnetic card to secure access to such information [6], but that has led to falsification of, and card and/or password theft. Additionally, there is the challenge that users need to remember a variety of passwords and pin numbers. This can be overcome by using e.g. biometric authentication which is based on human physical characteristics, for example: finger prints, face patterns, voice, handwriting, etc. The main advantage of voice biometry is that it can be used more easily in a variety of applications such as controlling devices at home or online transactions [7]. However, recognizing voices requires one to be able to identify unique patterns/features of a voice and then compare that with previously learned data, as shown in Figure 1.

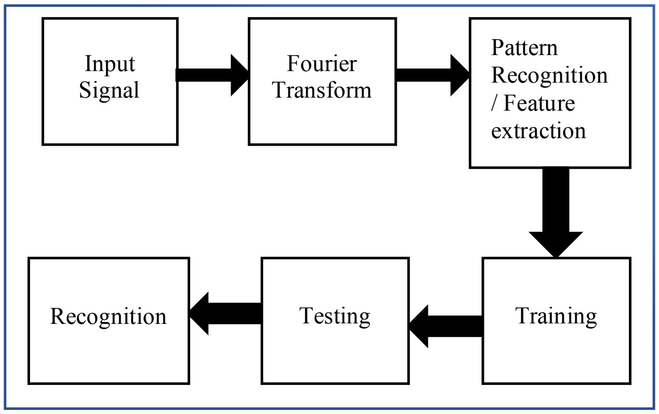

Figure 1: System Block Diagram

2. BACK GROUND

Speaker and speech recognition imply several challenges, which are linked in certain ways, making it challenging to keep a clear distinction between both. Similar challenges can be experienced if one was to purely use voice as a biometric identification, as many factors can affect the voice and therefore influence fraud prevention.

2.1 Speech Recognition

Speech recognition systems convert speech to text (see Figure 2) and thereby improve on the interface between humans and machines. Speech recognition is language dependent and aims to recognize what was spoken independent of e.g. "accents". The physical body generating the speech should not influence the recognition of words, which stands in contrast with voice recognition [8]. So, for speech recognition most development has aimed at making the tools become speaker independent [9].

While researchers have increased the recognition accuracy, the overall accuracy remains limited, which is partially due to the fact that they use syllables as their basic recognition. The limitations of syllables, lies in the fact that, as groups of letters, they can have a similar pronunciation, but a quite different meaning, making recognition challenging. For example: thought, through and tough, use a common group of letters, sound the same, but the actual choice of which word to use requires e.g. contextual information. Consequently, it would help if one would be able to distinguish more details about the speech when similar words are pronounced by the same person. As by better understanding the voice, one may find subtle differences between those words as such, which could then help to improve recognition accuracy.

Figure 2: Speech Recognition

2.2 Voice Recognition

The voice is a longitudinal wave and varies, based on how one uses his/her vocal folds in combination with other physical characteristics [10]. The main aspects of sound can be characterized as follows: 1) the sound wave's frequency is experienced as pitch [5], which describes the number of vibrations of each individual component within the signal and measured in Hertz; 2) the volume of the sound determines how loud or soft the sound is; 3) the tone is a pattern characteristic to the device generating the sound as such. For example, each instrument has a different tone, which is represented in the harmonics that are present besides the fundamental frequency.

Men and women differ by the number of vibrations per second for their vocal cords, which leads to their voice being experienced as "higher" (i.e. soprano) or "lower" (i.e. bas) [11]. While the male vocal cords vibrate at 100-130 times per second, for women this range is 180–220 vibrations per second. The volume of the voice relates to how it is amplified within the body, for which the mouth, nose and throat cavities are used, but also the chest can play an important role as "sound box". Understanding these physical characteristics and how they affect the produced voice is of significant importance in identifying individuals, and therefore key towards using the voice for biometric identification.

Figure 3: Voice Recognition

2.3 Biometric Authentication

In order to improve security, one needs to be able to authenticate someone to ensure that a person is who he or she claims to be. Authentication in the context of voice recognition is generally split into the need to: 1) verify, namely check if someone is who they claim to be through comparison with previously stored data for that person (see Figure 4); whereas 2) identification aims to identify a person out of a larger crowd (see Figure 5) [12], by seeing whether that person's profile equals that of anyone in the database or not. Obviously, one can only verify or identify a person if the relevant information has been previously obtained. When considering current authentication methods, one needs to be cautious of the fact that passwords can be easy to guess, but can be changed, while biometric solutions may be harder to mimic, but can never be changed as they are linked with the person's physical body. So once they have been compromised there is a serious challenge for that person and/or the verification/identification system.

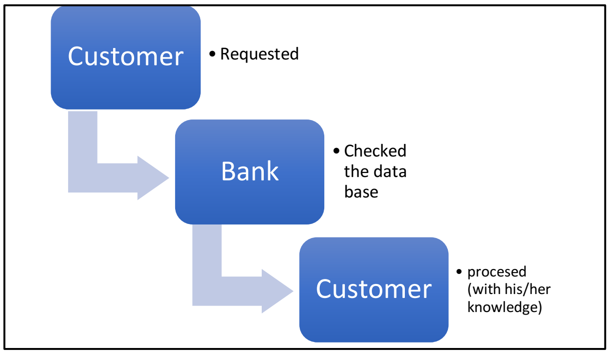

Figure 4: Speaker Verification

So, while biometric authentication aims to use a person's unique characteristics. As a technology it has gained popularity throughout the world as it can help reduce fraud cases [13] because it uses physical characteristics to identify oneself [6]. Biometric comes from the Greek, "bio" which means: "life" and "metric" which means: "to measure", and when combined they describe how to: "measure one's life". Biometric characteristics of individuals are unique [14] and can be used to authenticate one's identity for different systems. It is important to know that biometric technology has two types of properties which are either physical and/or behavioral. Physical properties include: DNA, fingerprints, facial recognition and iris/retina scan, while behavioral ones include voice, handwriting and signatures [15].

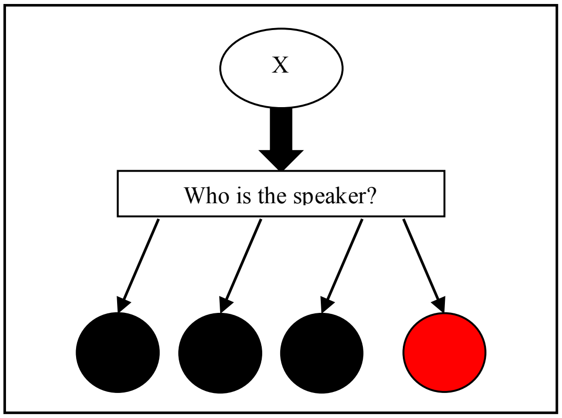

Figure 5: Speaker Identification

A biometric voice recognition system mainly focuses on identifying these unique characteristics and then store those within a database [11]. This requires one to understand which parameters are important to uniquely recognize one's voice, and considering these parameters are physical, but cannot necessarily be measured easily, one needs to identify how to identify these characteristics from the voice signal itself. This requires one to nearly work backwards in obtaining this information, as there may be characteristics that one would prefer to exclude from the profile to prevent the accuracy going down because one has e.g. a cold. The system should also be able to deal with challenges such as imitation, as in this situation the standard physical characteristics are "adapted on purpose" to create a different output, and it is yet to be seen how this could be identified appropriately.

Using voice profile data could also be useful for speech recognition tools if one can automatically detect who speaks and adjust the speech recognition profile for that person to improve overall accuracy.

3. PARAMETER IDENTIFICATION

Vibration of vocal cords and consequential production of voice have a variety of factors that in combination make the understanding of the actual production complicated. Before getting into more details of the actual production of vocal sound, it is useful to consider the human ear and how it processes sound, especially considering that voice recognition will need to work in direct combination with human hearing, and therefore any artificial system that needs to work similarly would benefit from working in a similar way. While there are various theories about the human ear, the most prominent ones either support a travelling wave or resonance theory [11].The travelling wave theory builds on the fact that a wave that entered the cochlear will be travelling continuously till it reaches a point where there is maximal displacement of the basilar membrane producing related spikes that go to the brain for decoding [10]. The main drawback of the travelling wave theory is the complication of accounting for the ear's delicate fine tuning [16]. On the other hand, the resonance theory is said to be the theory that is consistent with observations as well as being more elegant in explaining how we as humans hear. The theory allows for small energies to build up, cycle by cycle into an appreciable motion that triggers the tiny independently tuned hair cells in the cochlear.

While voice and speech recognition are in certain aspects interchangeable, they also have distinctive differences [17]. For example, the produced vocal sound depends on lip movements, shape of the mouth, vocal cords and resonance in throat, mount and nose cavities. The position of the jaw, tongue and the lips determines the resonance and anti-resonance of the vocal tract influencing the tone of the voice, ensuring each voice is unique [17]. However, other factors like size and shape of one's body also affect the voice as such.

The aim of speech recognition is not to depend on the physical characteristics of the body, because one wants to understand the speech to recognize the word, independent of who is speaking. One factor that affects speech recognition is language as such and most often also first language, if one e.g. speaks a foreign language, as accents tend to be related to the person's first language. Emotions also have a significant impact on speech. For example, if one is tired, then one tends to mumble, and not pronounce properly, which humans understand through combining contextual information and prediction to derive meaning, as demonstrated by our ability to finish the sentence of those you communicate with. However, current artificial systems are not able to derive context in the same way and/or benefit from prediction.

The existing speech recognition systems use syllables and hence researchers are also trying to use syllables to recognize voices. The main problem with using syllables for recognition is that in the case of speech the aim is to recognize words, which are formed be a group of syllables. Syllables are however a group of phonemes and so being able to recognize a voice requires one to hear the sound which is captured in the phonemes as those are most closely related to the physical profile of the human's vocal productive system. For example, when one says: "chef", which consists of three phonemes one should be able to use this information to recognize the person's voice profile.

In order to create a detailed voice profile, one needs to understand the tones of the voice, which requires an understanding of the various frequency components for which a Fourier transform can be applied to the individual input signals, providing a spectrum showing all frequency components and their respective "strength". By investigating this spectrum one should be able to identify specific patterns that are unique across different phonemes, which can then provide the features required to identify a voice. The machine should then be able to recognize voices based on their respective feature pattern.

These features would be an indirect representation of the human's vocal system including nose, mouth and throat cavities. These cavities amplified certain signals while damping other parts of the vibrations generated by the vocal folds, which makes one's voice unique. Each person has a unique voice, similar to fingerprints and mainly these differences in the resonant cavity are at the source of this. The mouth, throat and noise are the only resonators that shape the voice directly, because the sound wave passes through them on its way out. By adjusting those cavities through the use of muscles, one can then start to imitate other voices.

Additionally, there are also other physical characteristics that influence the voice, such as our body shape, muscles and the larynx. These other body parts contribute to producing less significant differences in the voice.

4. CONCLUSION

To improve the interface between human and machine, one needs to be able to identify the parameters that have an impact on voice/speech. Understanding these parameters and how important they are will help to improve the natural processing of human speech/voice to improve the artificial systems. It is also important to be able to differentiate between the parameters that relate to either voice and/or speech. The understanding of these parameters for voice recognition is particularly important with regards to using the voice as a biometric identification system to improve security. However, it is expected that by building up and understanding these voice related parameters that one will be able to improve speech recognition systems as such.

5. FUTURE WORK

The proposed method of performing frequency analysis through Fourier transformations will be implemented to the data gathered from a group of people speaking predefined sentences that cover the main phonemes. This data will be used to identify the patterns typical to each voice and so allow for human voice verification. However, one should also consider that users may try to imitate one's voice for several reasons. For instance in order to identify themselves to do online bank transactions on accounts different from their own.

6. REFERENCES

[1] A. O. Bayer and G. Riccardi "Semantic Language Models for Automatic Speech Recognition" *Signals and Interactive Systems Lab - University of Trento*, Italy, pp. 7–12, 2014, DOI= 10.1109/SLT.2014.7078541.

[2] L. Deng and X. Li, "Machine Learning Paradigms for Speech Recognition", *IEEE Trans. Audio, Speech Lang. Process.*, vol. 21, no. 5, pp. 1–30, 2013, DOI= 10.1109/TASL.2013.2244083.

[3] D. Guiming, W. Xia, W. Guangyan, Z. Yan, and L. Dan, "Speech Recognition Based on Convolutional Neural Networks," *International Conference on Signal and Image Processing,* pp. 708–711, 2016, DOI= 10.1109/SIPROCESS.2016.7888355.

[4] B. Soewito, "Smart Mobile Attendance System Using Voice Recognition and Fingerprint on Smartphone," *International Seminar on Intelligent Technology and Its Applications,* pp. 175–180, 2016,DOI= 10.1109/ISITIA.2016.7828654.

[5] T. Nakagawa, I. Nakanishi, Y. Itoh, and Y. Fukui, "Multi-modal biometrics authentication using on-line signature and voice pitch," *2006 Int. Symp. Intell. Signal Process. Commun. ISPACS'06*, pp. 399–402, 2007, DOI= 10.1109/ISPACS.2006.364913.

[6] W. Astuti and E. Byan Wahyu Riyandwita, "Intelligent Automatic Starting Engine Based on Voice Recognition," *IEEE SCOReD 2016*, 2016, DOI= 10.1109/SCORED.2016.7810061.

[7] S. Nafisah, "Evaluating the Usage of Short-time Energy on Voice Biometrics System for Cerebral Palsy", *International Conference on Informaiton Technology and Electrical Engineering (ICITEE),* pp. 1-6, 2016, DOI= 10.1109/ICITEED.2016.7863303.

[8] J. M. Baker *et al.*, "Research developments and directions in speech recognition and understanding, Part 1," *IEEE Signal Process. Mag.*, vol. 26, no. 3, pp. 75–80, 2009, DOI= 10.1109/MSP.2009.932166.

[9] V. Gupta, J. Bryan, J. Gowdy, "A Speaker-Independent Speech-Recognition System Based on Linear Prediction", *IEEE Transactions on Acoustics, Speech and Signal Processing,* Vol. 26, Issue 1, pp. 27-33, 1978, DOI= 10.1109/TASSP.1978.1163054.

[10] R. Ghaffari, A. J. Aranyosi, and D. M. Freeman, "Longitudinally propagating traveling waves of the mammalian tectorial membrane.," *Proceedings of the National Academy of Sciences of the U.S.A.*, vol. 104, no. 42, pp. 16510–5, 2007, DOI= 10.1073/pnas.0703665104.

[11] R. Singh, J. Keshet, D. Gencaga, B. Raj, "The Relationship of Voice Onset Time and Voice Offset Time To Physical Age Language", *International Conference on Acoustics, Speech and Signal Processing*, pp. 5390–5394, 2016, DOI= 10.1109/ICASSP.2016.7472707.

[12] R. Kannavara, G. Bebis, and N. Bourbakis, "An FPGA implementation of the local global graph-based voice biometric authentication scheme," *DSP 2009 16th Int. Conf. Digit. Signal Process. Proc.*, 2009, DOI= 10.1109/ICDSP.2009.5201175.

[13] S. Safavi, H. Gan, I. Mporas, and R. Sotudeh, "Fraud Detection in Voice-based Identity Authentication Applications and Services", *International Conference on Data Mining Workshops,* pp. 1074–1081, 2016, DOI= 10.1109/ICDMW.2016.115.

[14] P. Baraki, V. Ramaswamy, "Biometric Authentication of a User using Online Dynamic Signature," *Internaional Conference on Applied and Theoretical Computing and Communication Technology,* pp. 576–581, 2016, DOI= 10.1109/ICATCCT.2016.7912066.

[15] P. Korshunov and S. Marcel, "Impact of score fusion on voice biometrics and presentation attack detection in cross-database evaluations," *Journal of Selected Topics in Signal Processing*, vol. 4553, no. c, pp. 1–1, 2017, DOI= 10.1109/JSTSP.2017.2692389.

[16] W. He, A. Fridberger, E. Porsov, K. Grosh, and T. Ren, "Reverse wave propagation in the cochlea", *Proceedings of the National Academy of Sciences of the U.S.A.*, vol. 105, no. 7, pp. 2729–33, 2008, DOI= 10.1073/pnas.0708103105.

[17] G. Frewat, C. Baroud, R. Sammour, A. Kassem, and M. Hamad, "Android voice recognition application with multi speaker feature," *Proceedings of the 18th Mediterranean Electrotechnical Conference (MELECON)*, no. April, pp. 18–20, 2016, DOI= 10.1109/MELCON.2016.7495395.