v. 3, n. 5, 2023

Journal of
Engineering
Research

SUSTAINED VOWEL RECOGNITOR FOR NORMAL AND DYSPHONIC SPEECH

Mariana Regina Aguiar Catete
Department of Electrotechnics (DAELT) -
Universidade Tecnológica Federal do Paraná
(UTFPR)
Curitiba - Paraná
http://lattes.cnpq.br/2078048460696420

Marcelo de Oliveira Rosa
Department of Electrotechnics -
Universidade Tecnológica Federal do Paraná
Curitiba - Paraná
http://lattes.cnpq.br/0897919842779594

Abstract: This work aims to use speech recognition technology, through a probabilistic evaluation that the tool performs when analyzing the variation in the hit rate and accuracy of sustained phonemes of interest and variations inherent to the tool used. The focus was given to cases of people from 45 to 60 years old, in groups of individuals with normal speech and individuals with pathological speech, which are used to train and test the recognition system. Manual labeling was performed for the entire set of signals using the *Praat* tool; and speech recognition was performed using *Hidden Markov Models* (Hidden *Markov Model* or HMM) from the HTK tool (Hidden Markov Model Toolkit). The acquired results were a hit rate of 74.63% and accuracy of 39.18%. Therefore, even with low results, the method is effective and it is possible to optimize the method when training with a larger set of signals.

Keywords: HTK Toolkit; Hidden Markov Models; speech recognition.

INTRODUCTION

Around the 1930s, the first speech analysis and synthesis model was created in conjunction with the launch of the telephone industry, that is, for the telephone to achieve its purpose. Bell lab researchers also invested time in technologies for speech recognition (IACOMINI et al).

The process called speech recognition consists of mapping the speech signal into text, allowing its use to control actions in response to spoken commands, for example. One of the objectives of this technology is to make everyday tasks more practical, providing, among other advantages, more speed and productivity in situations in which the computer is useful to people whose hands or vision are otherwise occupied (SANTOS, 2008, 78).

In addition to communication, speech carries information about the anatomical and physiological conditions of the system that produces it (lungs, larynx and supraglottal tract). The conventional clinical procedure for its evaluation consists of the use of devices that allow the specialist to visualize such structures. Alternatively, the technique called acoustic analysis analyzes the speech signal, identifying anomalous behaviors that relate to diseases that affect such organs. One of them is the analysis of the vibration of the vocal cords (SHERWOOD, 2010; FOX, 2007; CRISTÓFARO SILVA, 2010).

The analysis of the vibration of the vocal cords verifies whether the sustained phonemes produce a quasi-periodic signal, with a well-defined harmonic structure and low amount of noise (CORDEIRO, 2016). A traditional way is to identify the fundamental frequency (or *pitch*) curve, which is always present in segments of the speech signal in which there is vibration of the vocal cords. As there are dysphonia in which this vibration is irregular or even non-existent (such as whispered speech, for example), the acoustic analysis is compromised.

To circumvent this problem and develop an automated acoustic analysis procedure, it is necessary to identify these phonemes (generally vowels) along the speech signal, even in situations of significant dysphonia - in which the vibration of the vocal cords is severely affected - for the application of analysis algorithms for present (or absent) harmonic content. In this work, a speech recognizer will be implemented to identify the sustained vowels of interest, allowing to locate them in the speaker's speech signal, since the identification of the phonemes will be strongly based on the so-called formant structure, derived from the geometry of the supraglottal structure.

GOAL

The objective of this work is to implement a speech recognizer to identify the segments in which there are sustained phonemes (mainly vowels) in speech signals produced by people with or without dysphonia, for further acoustic analysis.

For such implementation, a bank of speech signals from women with Reinke's Edema and others without any dysphonia was used. As it was not labeled (application of manual markings at the beginning and end of the phonemes), this work had the secondary objective of producing such labels, which could be used in future works.

MATERIALS AND METHODS

MATERIALS

For the development of the work, a bank of speech signals kindly provided by prof. Dr. Ana Paula Dassie-Leite from UNICENTRO (Irati campus) and duly approved by the ethics committee of that institution was assigned. It contains recordings of counting numbers (1 to 10) and vocalization of the phoneme /e/ for a few seconds (according to the person's phonatory capacity), of 30 women without dysphonia and 30 women with Reinke's Edema (dysphonia that is characterized by increased mass of the vocal cords, "thickening" the patient's voice) from 45 to 60 years.

First, the signals had their sampling frequencies adjusted to 16 kHz as required by the speech recognition system. This process was carried out using free *Audacity* software for convenience.

Subsequently, these signals were manually labeled (a process to identify sections - initial and final samples - of interest and assign a symbol or section for future use). In this case, the labeling identified the sustained phonemes: o (ô), oh (ó), in (nasalized i), eh (é), a (a), u (u), oi (oi), e (ê), i (i), ua (ua), C (consonants), N (speech out of data), sil (silence). For this time-consuming process, Praat was used. Figure 1 shows some examples of labeling using this *software*.

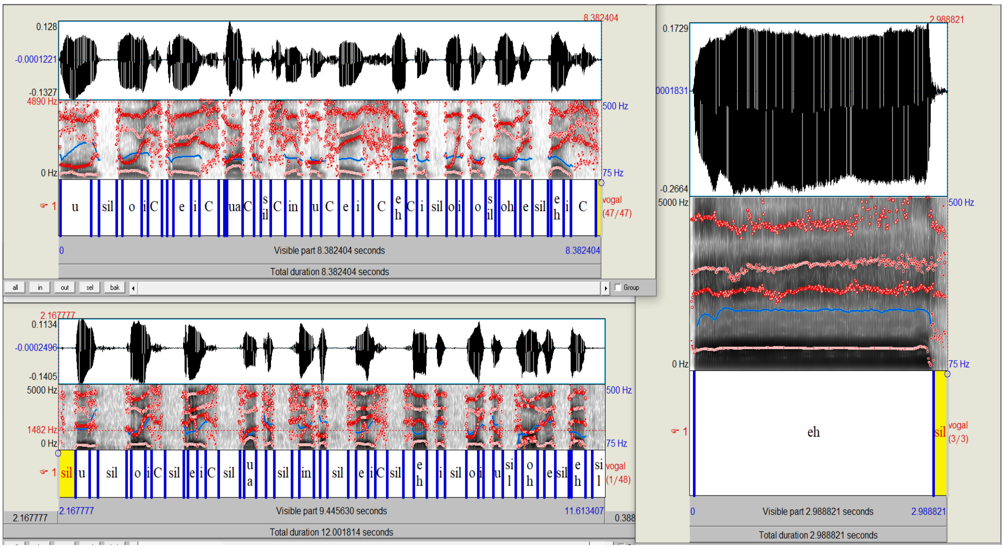
0.128
-0.0001221
-0.1327
4890 Hz
0 Hz
8.382404
500 Hz
75 Hz
u sil o i C e i C ua C sil C in u C e i C eh C i sil o i o sil oh e sil eh i C
vogal
(47/47)
0
Visible part 8.382404 seconds
8.382404
Total duration 8.382404 seconds
all in out sel bak
Group
2.167777
0.1134
-0.0002496
-0.1405
5000 Hz
1482 Hz
0 Hz
500 Hz
75 Hz
sil u sil o i C sil e i C sil ua sil in sil e i C sil eh i sil o i u sil oh e sil eh sil
vogal
(1/48)
2.167777
2.167777
Visible part 9.445630 seconds
11.613407
0.388
Total duration 12.001814 seconds
0.1729
2.988821
0001831
-0.2664
5000 Hz
500 Hz
0 Hz
75 Hz
eh
sil
vogal
(3/3)
0
Visible part 2.988821 seconds
2.988821
Total duration 2.988821 seconds

Figure 1 – examples of labeling.

Source: Own authorship (2021). Each tag is divided into three parts: the first part represents speech in the time domain; the second part represents speech in the frequency domain, the shades of black and white represent the frequency, the dotted lines in red represent the four speech formants and in blue the initial frequency of speech is represented, that is, the *pitch*.

For the recognition, the HTK tool (YOUNG et al, 2009) was used, which consists of a command line-based environment, and which was executed in the Linux environment (Ubuntu 18.04.4), in a virtual machine (for convenience). Such a tool employs a pattern recognition model based on a hidden Markov chain(or *Hidden Markov Model*), to transcribe acoustic sequences into sequences of symbols, starting from a training step. The tool also converts the speech signal into a spectro-temporal representation encoded in MFCCs (or mel-cepstral coefficients). All configuration of the models was done from text files that fed the HTK.

METHODS

The importance that speech recognizers have reached today is undeniable. For this reason, for the database to become complete, it is necessary to label the vowels using the Praat program. This tagging is necessary so that the formulated scripts have a learning base to be compared with the test base and, in the future, with any audio.

Audacity will be used in order to modify the frequencies of the data collected by Praat to the frequency and size at which the HTK works.

HTK is a program used to encode and decode audio. In order for it to work optimally, the audio is parameterized in small vectors that will facilitate the recognition of the isolated phoneme. What enabled a high performance training through the techniques inherent to the program (YOUNG et al, 2009).

RESULTS AND DISCUSSIONS

The Hidden Markov Model (HMM) is often used in modeling speech recognizers because it is a probabilistic model whose purpose is to perform signal decoding, as it produces good results as manipulators of statistical aspects and sequences of encoded sound information (YOUNG et al, 2009).

For example, a phrase consisting only of the word 'four'. Together with the initial and final silences, the transcription for this phrase would be:

sil C ua C o sil

We then have 4 acoustic sub-units to be trained, and as each one is modeled by a 3-state HMM (figure 2), we have a total of 12 probability density functions to be estimated. Assuming that this phrase was parameterized with 120 frames, we would have 10 frames for each state.

The first 10 symbols will initialize the first state of the first acoustic subunit - the first state of silence (sil) in the example -, the next 10, the second, and so on.

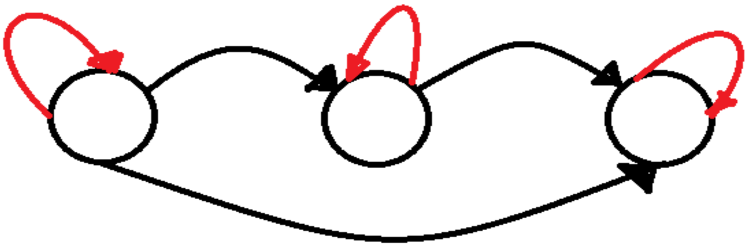

Figure 2 – 3-state right-left model.

Source: Own authorship (2021).

The initialization is done by simple counting: it is verified how many times each of the symbols occurred in these 10 frames, updating the counts of these symbols in the corresponding discrete probability density functions.

It is interesting to note that in the given example, there are two examples of the phonemes [sil], [u] and [a]. In this case, the counts of each of them is accumulated in the same function. Similarly, if we have more training phrases, the phoneme counts will be accumulated in the same function. Finally, these counts will turn into probability measures.

The results obtained with the database used for reduced training were: a hit rate of 74.63% with an accuracy of 39.18%, that is, of the total words, only 74.63% of the phonemes were recognized correctly. And, as the accuracy was low, it has been demonstrated that the method has a high error rate, which must have occurred due to the small amount of signals used for training.

CONCLUSIONS

Therefore, even with low results obtained, the method is effective. And, to improve the results, increasing the data used for the training stage will cause more phonemes to be recognized correctly, which will lead to a decrease in error rates and an increase in accuracy. A reassessment of labeling may also be considered to improve system performance.

ACKNOWLEDGMENTS

I would like to thank ''Universidade Tecnológica Federal do Paraná", Prof. doctor Marcelo Rosa for the opportunity to participate in a scientific initiation project, in addition to guiding, evaluating, making corrections and suggestions regarding the work and Prof. doctor Gustavo Nishida who made time available to evaluate and provide suggestions on how to improve this work and how the present work was carried out with the support of Fundação Araucária FA - Paraná/ Brasil, I extend my thanks to Fundação Araucária.

CONFLICT OF INTEREST

There is no conflict of interest.

REFERENCES

CORDEIRO, Hugo Tito. ***Reconhecimento de Patologias da Voz usando Técnicas de Processamento da Fala***. 2016. Disponível em: https://run.unl.pt/bitstream/10362/19915/1/Cordeiro_2016.pdf. Acesso em: 20 jan. 2022.

CRISTÓFARO SILVA, Thaís. ***Fonética e Fonologia do Português****: Roteiro de Estudos e Guia de Exercícios*. 2010. Editora Contexto.

FOX, Stuart I. ***Fisiologia Humana***. 2007. Editora Manole, Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788520449905/. Acesso em: 15 mai. 2022.

IACOMINI, Franco, *et al*. ***História das comunicações e das telecomunicações***. Disponível em: http://www.poli-integra.poli.usp.br/library/pdfs/5384658fb5ac96a9aa01791d9f6ff877.pdf. Acesso em: 15 mai. 2022.

SANTOS, Emilson Moreira dos. ***Engenharia Lingüística****: Uma tecnologia para apoiar as decisões gerenciais na era da Internet*. 2008. Editora E-papers, p 78. Acesso em: 15 ago. 2022.

SHERWOOD, Lauralee. ***Fisiologia humana****: Das células aos sistemas* . 2010. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788522126484/. Acesso em: 15 mar. 2022.

YOUNG, Steve, et al. ***HTK Book***. 2009. Cengage Learning Brasil, 2021. Disponível em: https://htk.eng.cam.ac.uk/download.shtml. Acesso em: 20 out. 2021.