Development of a voice database to aid children with hearing impairments

2011 J. Phys.: Conf. Ser. 332 012047

(http://iopscience.iop.org/1742-6596/332/1/012047)

Development of a voice database to aid children with hearing impairments

M G Kuzman[1], P D Agüero[1], J C Tulli[1], E L Gonzalez[1], A J Uriz[2] and M P Cervellini[1]

[1] Laboratorio de Comunicaciones, Facultad de Ingeniería, Universidad Nacional de Mar del Plata, Juan B. Justo 4302, Mar del Plata, Argentina
[2] CONICET - Laboratorio de Comunicaciones, Facultad de Ingeniería, Universidad Nacional de Mar del Plata, Juan B. Justo 4302, Mar del Plata, Argentina

E-mail: {melisakuzman,pdaguero}@fi.mdp.edu.ar

Abstract.
In the development of software for voice analysis or training, for people with hearing impairments, a database having sounds of properly pronounced words is of paramount importance. This paper shows the advantage that will be obtained from getting an own voice database, rather than using those coming from other countries, even having the same language, in the development of speech training software aimed to people with hearing impairments. This database will be used by software developers at the School of Engineering of Mar del Plata National University.

1. Introduction

Hearing impairments approximately correspond to 18 % of all disabilities in Argentina. This statistic has been published by the CAS (Argentinan Federation of Deafs) [1], which keeps records of thousands of people with hearing impairments.

Early childhood is a particularly critical period for language acquisition and children development. In general, hearing losses cause difficulties for social adaptation [2]. Therefore, it is important to increase the opportunities of communication for preschool children with hearing losses, who have poor pronunciation skills due to the lack of auditory feedback.

Hence, it is essential to assist them as soon as possible in order to palliate this situation. For this purpose, creating properly designed applications for computers is very important.

Currently there are several programs that are used for oral speech training and rehabilitation of people with hearing impairments. Among them, "Dr. Speech"[3], which is a software for voice training. Another development is carried out at the University of Zaragoza. It is named "Alborada 13-A"[4],it is a software application for recognition and evaluation of the voice, which is able to detect pronunciation errors. This application is designed to help people with severe hearing losses in the speech training.

These applications have been created with the aim of improving communication capabilities, contributing in this way to the social interaction of people with hearing impairments

Software for voice recognition and evaluation is generally formed by three main components: a **voice database** to estimate statistical models or templates, a **voice recognition engine**,

and finally, a user interface that allows for an **evaluation** of the utterance or gives users a visual feedback of their pronunciations.

A critical aspect of speech databases used at the laboratory of this research group at Mar del Plata National University, is the Spanish origin. This fact brings problems among acoustic analysis, since the linguistic forms of some words and their corresponding pronunciations often significantly differ with respect to the Argentinian dialect.

Considering that, the first goal that arises is the building up of a local database, it means the argentinian dialect. Its construction is divided into two groups: the first contains information about children with hearing impairments, while the second group contains voices of children with normal hearing abilities.

Once the voice database is done, it is intended to be used to train the SPHINX recognition engine [5] (a voice recognition tool written for Java). This application makes possible the evaluation of deletions, insertions, or substitutions of phonemes, a clinically phenomenon known as dysarthria [2]. Finally, the data and the recognition engine will be used to develop programs for people with severe hearing losses, such as SPO [6].

The paper is organized as follows. Section 2 describes several possibilities for data voice acquisition and analysis. Section 3 shows the achieved diphoneme coverage for the Spanish voice database, and how it could be improved. Finally, Section 4 details the conclusions of the study, analyzing future issues on the Spanish database.

2. Methods

In order to develop software projects applications that help people with severe hearing losses in their oralization practices, it is necessary to use a voice recognition system [7], also known as **ASR**[1]. It is a computer tool used to process voice signals and produce the transcription of the content. Because the ASR system uses statistical models, it is necessary to obtain appropriate amounts of data in order to get precision in the acoustic models.

2.1. Selection of the data to record

Depending on the application, there are several possible groups of phonemes that may be included in any voice database, which are listed below:

- Phonemes: it is the minimum unit used in voice recognition systems[2]. Spanish language has 18 consonants and 5 vocals [8]. In this paper work an inventory of 24 phonemes and 7 allophones is used, which are defined by Alarcos (1950) [9]. Because of the contextual variations, these units must be recorded in the context of words or sentences.
- Diphonemes: these units consider the coarticulation between speech sounds, by taking into account the left or the right context. There are 961 different combinations, but not all of them exist in Spanish.
- Triphonemes: these units take into account the central phoneme and the left and right context. In this case there are 29791 combinations, a higher amount than the previously case. However, many combinations must be deleted because they are not used in Spanish.
- Words and sentences: these are the only units with semantic meaning. Nevertheless, they are not useful to be assembled for recognizing any word.

For the development of the voice database, words and sentences are going to be chosen according to their diphoneme coverage. These units give flexibility to the system, and require a smaller number of words to be recorded. In order to have a good coverage of Spanish words,

[1] Stands for Automatic Speech Recognition.
[2] The phoneme is the minimum unit without semantic meaning in the spoken language.

the recorded data will be carefully selected to minimize resource needs (mainly time, people and storage). The resulting database will be useful to develop different aid applications.

2.2. Recording conditions

Recordings will be performed with open-source software tools capable of handling and manipulating sounds. The capabilities that are of interest in this work are: sound recording, speech analysis and phonetic annotation.

The data recording is going to be made with a dynamic microphone (SPK Pro-92 XRL, unidirectional, with a frequency response between 50- 16000Hz, output impedance of 600 Ohms and a sensibility of -52dBV $\pm$ 3dB).

All recordings must be done with the supervision of a voice therapist in order to help children with their utterances. The recording will be performed with the support of a professional in order to instruct the children. It is thought to generate an appropriate environment for the children's performance, and it will give naturalness to the speech [2].

Voice database recordings must take into account several specifications. Some of them are:

- Signal-to-Noise Ratio (SNR).
- Sampling frequency.
- Number of quantification bits.
- Reverberation time.
- Dynamic range.
- Presence of saturations (mostly in plosive sounds).

In the next paragraphs each specification is going to be explained, and the strategies to improve them will be described.

Signal-to-Noise Ratio. The recording room must be of appropriate size and with soundproof walls (irregularities on the walls, curtains, etc.) in order to avoid inconvenient acoustics reflections. To improve the **SNR**, recordings will be done with a laptop working on battery mode (without connection to the electric network), to avoid low frequency electric interferences. The speaker will be far from external noise like cellphones, air conditioning or computer's fans (including the laptop computer). The SNR is a very important specification of a system. It provides the relationship between the power of the signal under study and the noise power from other sources.

Sampling frequency and number of sampling bits. This parameter must be appropriate according to the required data quality. Sampling frequencies are limited by the capabilities of the computer's sound card . The sampling frequency of 16 Khz is chosen following the specifications of the project "Vocaliza". [4].

Reverberation time. RT is defined as the time it takes for the signal to fall 60 dB, since the speaker finished the phonation. Thus, it is considered that reflections finish when the intensity of the perceived signal correspond to 0,1% of the original signal. Minimization of this parameter results in a more accurate information obtained from the data.

Use of dynamic range. When the voice is recorded, some sounds may be louder than others, for example plosive sounds. These are produced when the vocal tract is closed, stopping all airflow and pressure is built up behind the occlusion. If this air is suddenly released, for example when pronouncing the occlusive /p/, the abrupt change of air pressure may cause a saturation in the

recording. A good compromise between dynamic range and signal level is necessary in order to produce reliable recordings.

Voice recognition engine
The goal of the automatic speech recognition system is to produce a word sequence (or labels) which are the textual representation of the acoustic signal. This task is not trivial, because the acoustic signal not only contains words, pauses, but also disfluences , environmental noise, articulatory noises (lips, breathing), etc.

In order to perform this task, the automatic recognition system uses tools of statistical modeling and sequence decoding. It applies a set of simplifications in the approach to get an implementable solution.

The limitation of the vocabulary is the first simplification made by ASR systems. Its size could be a few words (e.g. recognition system of verbal orders) or tens of thousands of words (large vocabulary recognition system). It is not possible for the ASR to recognize words that are not in its dictionary, because it will not find the boundaries of them. For example, the sentence *"la casa de la pradera está habitada por un ermitaño"* could be pronounced *"lacasadelapradera estáhabitada porunermitaño"*. As seen, there are pauses before the words *pradera* and *habitada*, but the others are pronounced without any silence between words. Therefore, it is necessary to know the words to find their boundaries.

Nowadays, automatic speech recognition systems made their task based on statistical models of the acoustic signal (acoustic modeling of the phonemes) and the language (language modeling using words) [10]. The parameters of the statistical models are estimated using collected data, such as acoustic signals, and their corresponding phonetic and orthographic transcriptions. The reliability of the statistical models will be bounded by the amount of available corpus.

In the development of software for hearing disabled people, the functioning of the automatic speech recognition technology is partially modified to detect mispronunciations. In this case, the ASR does not have to recognize the word, but it should analyze the word that the speaker was asked to pronounce. In these conditions, it is possible to improve the analysis capabilities of the system incorporating possible alternative pronunciations due to dysarthria.

To be consistent with the use of JAVA in the development of systems to aid people with hearing impairments in their oralization practices, the ASR SPHINX was chosen to segment the recorded voice into phonemes to perform the analysis.

SPHINX-4 is an updated recognition system which was programmed in JAVA language. This system was created by the joint collaboration of the Carnegie Mellon University, Sun Microsystems Laboratories, Mitsubishi Electric Research Labs (MERL) and Hewlett Packard (HP), with the contributions of the University of California of Santa Cruz (UCSC) and the Massachusetts Institute of Technology (MIT).

3. Experiments
In order to know the word coverage of phonemes of the Alborada's database, the automatic phonetic transcription of the 57 recorded words was performed. It was done by a software that converts the word into the canonical phonetic transcription. Then, the phonetic transcription was used to get the phonemes of the words. The results show that the Alborada's database contains 147 phonemes. Figure 1 shows the phonetic transcription of a group of words made by the previously mentioned software.

The analysis of the coverage of the Spanish words with some specific phonemes was made using the written CREA corpus. The reference corpus of Spanish is a set of texts of different sources, which are stored in a computer support. The last one has the information of the words, their meanings and their context [11]. The Real Academy of the Spanish Language explains

Table 1. Words split into phonemes

Words	Phonemes
boca	_+b b+o o+k k+a a+_
bruja	_+b b+r r+u u+x x+a a+_
cabra	_+k k+a a+B B+r r+a a+_
campana	_+k k+a a+m m+p p+a a+n n+a a+_
casa	_+k k+a a+s s+a a+_
caramelo	_+k k+a a+r r+a a+m m+e e+l l+o o+_
casa	_+k k+a a+s s+a a+_
clavo	_+k k+l l+a a+B B+o o+_
cuchara	_+k k+u u+tS tS+a a+r r+a a+_
dedo	_+d d+e e+D D+o o+_
ducha	_+d d+u u+tS tS+a a+_
escoba	_+e e+s s+k k+o o+B B+a a+_
flan	_+f f+l l+a a+n n+_
fresa	_+f f+r r+e e+s s+a a+_
fuma	_+f f+u u+m m+a a+_
gafas	_+g g+a a+f f+a a+s s+_
globo	_+g g+l l+o o+B B+o o+_
gorro	_+g g+o o+rr rr+o o+_
grifo	_+g g+r r+i i+f f+o o+_
indio	_+i i+n n+d d+j j+o o+_
jarra	_+x x+a a+rr rr+a a+_
jaula	_+x x+a a+w w+l l+a a+_
lápiz	_+l l+a a+p p+i i+T T+_

that the corpus is a representative state of nowadays Spanish, because the text it contains has been selected according to the current parameters.

The 1000 words most used in Spanish according to CREA have been transcribed into diphonemes. The results show that there are 440 phonemes in the CREA Corpus. Therefore, the 147 diphonemes of the Alborada's database are not enough to cover the group of 1000 most frequent words.

In order to get a list of words that cover the 440 phonemes, a greedy algorithm was used. It selects a set of words to minimize the number of words needed to cover the 440 diphonemes. The final total was 320 words.

Applying the same procedure, it was possible to find the number of triphonemes necessary to cover the CREA Corpus. The results indicate that it is necessary 1778 words to comprehend the 2919 triphonemes in the Corpus.

Therefore, the use of diphonemes implies a positive trade-off between the reliability of statistical models and the amount of data that is necessary to record to get a good coverage.

4. Conclusion

In this paper has been mentioned that people with hearing impairments may experiment some problems when social interaction is required. Also, there are mentioned some applications that are useful for these people, in order to improve their speech abilities using specific softwares. Besides, it was emphasized the importance of having a suitable voice database for

the development of these applications.

The origin of the voice database used at the laboratories of the School of Engineering is Iberic Spanish. The main problems are the different linguistic forms compared with the Argentinian Spanish. An own voice database will give the chance of improving the applications developed by this research group, and made it freely available to the community, like the project "Sistema de Práctica de la Oralización" (SPO) [6, 12, 13, 14].

The goal of this project is, on one hand, to get an Argentinian voice database and improve the results of voice reeducation with respect to those obtained with the Iberic Spanish database. On the other hand, it is proposed a different method to select the phonetic content that will give flexibility to the voice recognition system for many existing or new applications.

The results of the experiments show that the Spanish database has not enough coverage of the 1000 most frequent Spanish words. In order to get that coverage, it is necessary to record at least 320 carefully chosen words. This gives the chance to create and develop new applications, using a reliable and versatile voice database.

References

[1] "http://www.cas.org.ar," Tech. Rep.
[2] R. L. Belloto, *Voz y pronunciación en el niño discapacitado auditivo.* Ediciones ARES, 1974.
[3] "http://www.drspeech.com," Tech. Rep.
[4] O. Saz, E. Lleida, C. Vaquero, and W.-R. Rodríguez, "The alborada-i3a corpus of disordered speech," Tech. Rep., 2010.
[5] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," CMU, Tech. Rep., 2004.
[6] J. M. Garin, P. D. Agüero, J. C. Tulli, E. L. Gonzalez, and A. J. Uriz, "Spo: una ayuda en java para la oralización de hipoacúsicos," Tech. Rep., 2010.
[7] H. van den Heuvel, K. Choukri, C. Gollan, A. Moreno, and D. Mostefa, "Tc-star: New language resources for asr and slt purposes," in *Proceedings LREC 2006*, 2006, pp. 2570–2573.
[8] J. Llisterri and J. B. Mariño, "Spanish adaptation of sampa and automatic phonetic transcription," Tech. Rep., 1993.
[9] E. Alarcos, *Fonología española.* Gredos, 1950.
[10] S.-C. Yin, R. Rose, O. Saz, and E. Lleida, "Verifying pronunciation accuracy from speakers with neuromuscular disorder," in *In INTERSPEECH-2008*, 2008, pp. 2218–2221.
[11] "http://www.rae.es," Tech. Rep.
[12] J. M. Garín, P. D. Agüero, J. C. Tulli, and E. L. Gonzalez, "Sistema de práctica de la oralización en plataforma java," in *SICA*, Noviembre 2009.
[13] J. M. Garin, P. D. Agüero, and J. C. Tulli, "Nuevos aportes al entrenamiento de personas hipoacúsicas sistema de entrenamiento para personas hipoacúsicas en plataforma java," in *TISE*, Diciembre 2008.
[14] J. M. Garin, P. D. Agüero, and J. C. Tulli, "Nuevos horizontes - aportes al entrenamiento de personas hipoacúsicas," in *SICA*, Septiembre 2008.