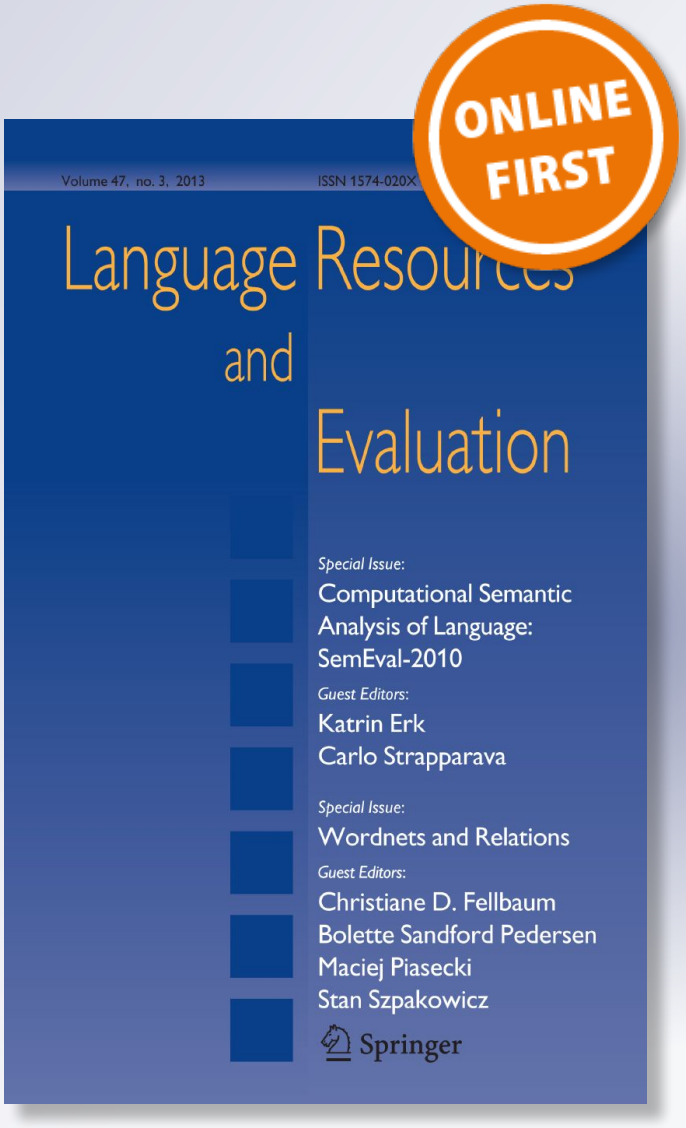
Emilia: a speech corpus for Argentine Spanish text to speech synthesis

Humberto M. Torres, Jorge A. Gurlekian, Diego A. Evin & Christian G. Cossio Mercado

Language Resources and Evaluation

ISSN 1574-020X

Lang Resources & Evaluation
DOI 10.1007/s10579-019-09447-7



Springer

Lang Resources & Evaluation
https://doi.org/10.1007/s10579-019-09447-7

ORIGINAL PAPER

Emilia: a speech corpus for Argentine Spanish text to speech synthesis

Humberto M. Torres[1] · **Jorge A. Gurlekian**[1] · **Diego A. Evin**[2] · **Christian G. Cossio Mercado**[3]

Abstract This paper introduces Emilia, a speech corpus created to build a female voice in Spanish spoken in Buenos Aires for the Aromo text-to-speech system. Aromo is a unit selection text-to-speech system, which employs diphones as units of synthesis. The key requirements and design criteria for Emilia were: to synthesize any text in Spanish into high-quality speech with a minimum corpus size. The text corpus was designed to guarantee the phonetic and prosodic coverage. A three-stage strategy was used: in the first stage, 741 sentences were designed with all of the syllables of Spanish spoken in Argentina, with and without stress, and in all positions within the word; in the second stage, 852 sentences were added to balance out the distribution of the diphones; and after a perceptual evaluation of the quality of synthesized speech, in the third and final stage, 625 sentences were added to achieve the specified unit coverage, and to introduce sentences with more complex syntactic and prosodic structures. Issues from all three corpus building stages are reported. The paper also presents the results from the quality perceptual evaluations of the

✉ Humberto M. Torres
hmtorres@conicet.gov.ar

Jorge A. Gurlekian
jag@fmed.uba.ar

Diego A. Evin
diegoevin@gmail.com

Christian G. Cossio Mercado
ccossio@dc.uba.ar

[1] Laboratorio de Investigaciones Sensoriales, INIGEM, CONICET-UBA, Av. Córdoba 2351, 9 Piso Sala 2. C.A.B.A. (1120), Buenos Aires, Argentina

[2] Center for Research and Transfer in Acoustics (CINTRA), UTN-FRC UA CONICET, Master M. López esq. Argentine Red Cross, University City, X5016ZAA Córdoba Capital, Argentina

[3] Departamento de Computación, FCEN, UBA, University City, C1428EGA Buenos Aires, Argentina

Published online: 02 February 2019

synthesized voice. Emilia has a duration of three hours and 15 minutes; its speech quality synthesized with Aromo system is similar to the level obtained with commercial systems, with a real-time ratio less than one.

Keywords Speech corpus design · Text-to-speech · Argentine Spanish · Phonetic corpus · Phonetic transcription

1 Introduction

Most of state-of-the-art high-quality text-to-speech (TTS) systems use unit selection and concatenation approach (King 2014). The basic idea behind this technique consists in having a database with multiple instances of each speech unit, and a feature vector associated with each one. Given a text input in run time, the TTS engine selects the most suitable unit sequence, according to a set of selection criteria. Proper design and construction of a speech corpus for building a voice has a strong impact on the TTS systems speech quality (Möbius 2000; Kurtic 2004; Lambert et al. 2007; Chevelu and Lolive 2015; Niebuhr and Michaud 2015). It has been shown that there are different factors that affect the quality of a corpus. These include, among others, the text of the sentences, the type of units, the accuracy of labeling, the number of instances of each concatenation unit and the corpus coverage.

A speech corpus is a collection of speech samples recorded by one or more speakers with a set of associated timestamp labels. These labels identify the linguistic content of the speech such as half-phones, phones, diphones, syllables, words, phrases, sentences, prosodic marks, part-of-speech tags, syntactic categories, among others (Llisterri 1999).

In unit selection TTS systems, the speech corpus is used to extract the inventory of speech units and build or train the models that predict prosodic attributes. In these systems, the diphone is the most common unit defined as the speech segment extending between the stable sections of two consecutive phones (Dutoit 1997).

An ideal database for these types of TTS systems will contain all possible variations for each unit, including phonetic and prosodic context variations. The number of units to consider is usually large, even for small-domain systems. It is impossible to build an optimal database for unrestricted domains, but in general it is accepted that the higher the number of instances of each concatenation unit, the higher the quality of the synthesized speech produced. This encourages the construction of large databases with many occurrences of each synthesis unit, containing different levels of segmentation and labelling (Kawai and Tsuzaki 2002).

The steps to build a corpus for a unit selection TTS system are: corpus specification; text corpus creation; speaker selection and recordings; speech corpus labelling; and corpus quality assessment. These five steps will be briefly described in the following sections.

1.1 Corpus specification

A clear idea about the *why*, *what* and *how* is fundamental when addressing the creation of a speech corpus. This should be translated into a set of specifications used as a corpus construction guideline and for its subsequent validation. The items that should be contemplated in corpus specifications include domain of application, type and number of concatenative units, corpus coverage, desired quality of synthesized speech, real-time requirements, hardware and software platform constraints, among others.

TTS systems can be classified as: domain restricted to an application, for example for use in a Global Positioning System device; multi-domain, such as TTS systems used in IVRs (Interactive Voice Response); or open domain for general purposes, such as an application to read a newspaper. For high-quality synthesis, both the text corpus and the speech styles in the recordings must be representative of the system domain, the so-called target domain (Alıas et al. 2003; Black and Lenzo 2000).

Different segments of speech have been proposed as the concatenation unit: half-phones, phones, diphones, triphones (Kishore and Black 2003; Grůber et al. 2007); polyphones (Hon et al. 1998; Torres and Gurlekian 2008); syllables, words (Lewis and Tatham 1999); and phrases or combinations of them (Prudon and d'Alessandro 2001; Boëffard 2001). In general, there are two conflicting characteristics that determine the choice of unit to be used: the units have to be long enough to capture and keep up the dynamics of coarticulations (Peterson et al. 1958); and on the other hand, a reduced number of units is preferred to reduce the search space during the unit selection process. This trade-off is the main reason why the diphones are the most commonly used units in speech synthesis by concatenation. Longer units can reduce the concatenation artifacts and improve the modeling of coarticulation dynamics but then a much larger number of units are needed to achieve the coverage required. Additionally, the language and the application domain of a TTS system also influence the choice of the units to be employed. For example, in the context of an IVR system, it is common to include courtesy phrases that are frequently used; they can be fixed a priori, and included as a single piece into the corpus, guaranteeing excellent synthesis quality.

An important metric to characterize a unit selection TTS system is the coverage ratio, which measures the probability of being able to synthesize a random sentence with the units available in the corpus (van Santen 1997). Considering the diphones as a synthesis unit and taking into account the 30 phones plus allophones for Argentine Spanish (Gurlekian et al. 2001a), in theory, 100% coverage can be achieved with just 961 diphones. However, the unit selection process takes into account not only the phonetic identity but also other features such as the stress condition, the phonetic context and the acoustic parameters, among others. Therefore, ensuring the coverage of a corpus is not a trivial task (van Santen 1997). The coverage problem is NP-hard, which is usually solved by greedy algorithms (Franois and Boëffard 2002; Bozkurt et al. 2003). Different factors can be taken into account when defining coverage of a corpus: synthesis units, prosodic features, and linguistic factors, among others. Some of these factors are orthogonal.

Furthermore, many of the speech features respond to large number of rare events (LNRE) distributions (Möbius 2003). Therefore, a speech corpus containing all possible variations of a language would be huge in size, quite possibly making it impossible to build this corpus. An alternative approach to solve this problem is to consider coverage hierarchies, for example, ensuring first a phonetic coverage; secondly, considering stress conditions; followed by the position of the units within the sentences, and so forth (Andersen and Hoequist 2003).

Initial databases used with the unit selection approach had a duration of a few minutes of speech (Campbell 1996). Currently it is broadly accepted that the database must have a duration of several hours in order to achieve high quality synthetic speech (Campbell 2005; Bonafonte et al. 2006; Umbert et al. 2006; Ni et al. 2007; Matoušek et al. 2008; Oliveira et al. 2008). The size of the corpus could be conditioned by several factors: the corpus coverage; the memory capacity, which could be a restriction in embedded applications; the speed required for the synthesis, among others. As a rule, larger corpora are more expensive to build, require more storage space and have high run times. In general, corpus size is tied to the desired coverage and to the method used to achieve it. Once coverage is achieved, some techniques can be applied to speed up the synthesizer, such as the application of pruning techniques (Lu et al. 2015). There are two approaches to prune the units set: top-down or bottom-top (Rutten et al. 2002; Krul et al. 2007; Bellegarda 2008; Grůber et al. 2014). In the top-down approach, the selection frequency of each unit in the corpus is estimated using a reference text, followed by the removal of less frequent units or even complete sentences from the corpus containing these units. In the bottom-top approach, the corpus is analyzed in order to remove redundant units. Strategies combining both approaches have also been proposed, for example, to apply a top-down pruning but considering the redundancy of the corpus (Karabetsos et al. 2009). Because the corpus pruning is done after its development (recording and labeling), and it involves discarding units or sentences, this can be considered a waste of energy. Furthermore, corpus pruning affects the quality of the synthesized speech (Hansakunbuntheung et al. 2005).

1.2 Text corpus creation

Text design for TTS speech corpus plays an important role in the synthesized speech quality. Regardless of methods and technologies used on a TTS system, if the corpus is not well built, the synthesized speech quality will be poor (Möbius 2000; Chalamandaris et al. 2011).

Corpus text domain should be as close as possible to the working environment of TTS system. This helps mitigate commitment between text, duration and coverage of the corpus.

There are different approaches for designing a text corpus. One alternative is to maintain the natural distribution of synthesis units over the application domain, which is usually done as a random sample of that domain. This does not ensure the coverage, but it has been shown to generate a more natural synthetic voice (Lambert et al. 2007). Another alternative is to ensure the corpus coverage at different levels, such as phonetic or prosodic, among others (Franois and Boëffard 2001). This

approach would make it possible to synthesize the majority of the likely input text. Several algorithms have been proposed to generate a set of optimal text from a large reference text based on some coverage criteria (van Santen 1997; Franois and Boëffard 2001; Franois and Boëffard 2002; Chevelu et al. 2008; Zhang et al. 2010).

Some authors prefer to start with a large amount of reference text, and then apply different techniques to reduce the corpus to the desired size (Franois and Boëffard 2002; Kelly et al. 2009; van Santen 1997). In contrast, others propose to start with a small text that fulfills certain specifications, for example, that all synthesis units must be present, and then keep adding more text to achieve the desired size (Breen and Jackson 1998). Matoušek and Psutka (2001) propose to create a corpus that follows the distribution of real speech, and they present an iterative algorithm for selecting a subset of sentences containing a minimum of occurrence for all concatenation units. One objection to this algorithm is that the number of sentences needed to achieve a minimum of occurrence for each unit can be extremely large and difficult to achieve. Other studies raise the question from where and how must be selected the text to be used (Lambert et al. 2007). Beutnagel and Conkie (1999) analyze the pruning of infrequent units, and they conclude that adding rare units to the corpus contributes significantly to the quality of the synthesized speech.

1.3 Speaker selection and recordings

Speakers selected to record a speech corpus must have some particular attributes: a pleasant voice, clear articulation, the ability to maintain the style of speech through different recording sessions. Extracted concatenation units must also generate high-quality synthetic speech (Syrdal et al. 1998; Coelho et al. 2009; Niebuhr and Michaud 2015).

In most cases, it is difficult to test whether a speaker meets these requirements. Furthermore, a speakers quality of natural speech of a speaker is not always an indicator of the quality level of his synthesized speech (Syrdal et al. 1998). One way to select the speaker is to create a mini-corpus with the voice of each speaker candidate, build the voice and perceptually evaluate the synthesized speech (Umbert et al. 2006). This approach has two cons: it is too expensive and the quality of the synthesized speech depends on the mini-corpus size.

Generally, recordings should be made in short-duration sessions spaced over close intervals to avoid speaker fatigue and reduce the risk of any variations in the manner of speaking. The speaker should understand the text to be read before recording sessions. Recordings should be of high quality, made in a suitable, noise-free acoustic chamber. It is customary to monitor the speech signal to detect any noise or inconsistencies in recordings.

1.4 Speech corpus labelling

Audio files must be labeled at various levels: phonetic, syllables, words, phrases, prosody, among others. These labels are used to extract concatenation units, train the models that predict prosodic features and estimate input features for the unit selection module.

Proper annotation of the corpus improves the quality of synthesized speech. Labeling may be manual, automatic or combined (Chu et al. 2006). A skilled workforce is needed to perform the labelling by hand, which is an extremely time-consuming and very expensive process. More than one labeler can be used to reduce the labeling time, but it gives rises to the need to ensure uniformity in the labeling criteria and inter-labeler agreement (Hoeckel 1989; Eisen 1993; Pitt et al. 2005; Bayerl and Paul 2011). Therefore, in general, automatic labeling is carried out, which may be followed by a manual, full or partial correction. In most cases, to assess the quality of automatic labeling, it is measured against a small subcorpus with a manual labelling. However, there are reports that indicate that perceptual measures of synthesized speech are the only valid form of measurement (Syrdal et al. 2000; Kawai and Toda 2004; Adell et al. 2005).

1.5 Corpus quality assessment

A validation process makes it possible to assess whether the corpus meets its specifications. This process involves the analysis and documentation of the corpus (van den Heuvel et al. 2008; Schiel et al. 2012). Validation can be done by the producer (in-house) or an accredited third-party entity (external) during the creating process of the corpus or after completion, with either manual or automatic tools. In-house validation and throughout the creation process allows producers to adjust processes to ensure compliance with the specifications. An external validation over the final version of the corpus makes the certification of the resources for its distribution easier.

Evaluation of a TTS speech corpus refers to the testing of the quality of synthetic speech generated with the corpus (Dybkjær et al. 2007). The result of the assessment depends on the TTS system and the corpus alike, and it is difficult to separate one from the other. One approach is to compare the performance of several corpora with the same TTS system. Assessment of synthetic speech takes into account two aspects: intelligibility and naturalness. Different perceptual methods have been proposed to measure these properties of synthetic speech, including: International Telecommunication Union (ITU) recommendations P.85. (1994) and P.800 (1996), Semantically Unpredictable Sentences (SUS) (Benoît et al. 1966), among others.

The standard Mean Opinion Score (MOS) (P.800 1996) is a general-purpose test. It provides average scores for natural and artificial speech. A fixed-point scale is used to evaluate overall quality. Common scales consist of five or ten points. These scales are: non-linear; vary from one listener to another; without absolutes; language dependent; and the scores obtained are only meaningful in the context in which they were obtained (P.85. 1990; Streijl et al. 2016; Watson et al. 2001). Furthermore, a low-quality stimulus is rated worse if the set of test stimuli is high quality, compared to a set of poor-quality test stimuli (Hall 2001; Taylor 2009).

Focusing on a particular aspect or dimension of speech is an alternative to assessing overall speech quality. ITU-T P.85 is a standard where several MOS are collected following this approach. But focusing on one dimension of speech can be very difficult for a listener. In addition, defining which speech dimensions should be

evaluated and ensuring that these dimensions are not correlated is still under study (Alvarez and Huckvale 2002; Hinterleitner et al. 2013; Hirst et al. 1998; Mayo et al. 2005; Sityaev et al. 2006; Taylor 2009; Vainio et al. 2002; Viswanathan and Viswanathan 2005).

A comparison MOS (CMOS) is the result of a comparison category ranking test. Two stimuli are presented and subjects have to choose which sample is better and by how much. Pairs of samples are played at random (P.800 1996).

Semantically Unpredictable Sentences test (Benoît et al. 1966) is focused on evaluating the intelligibility of systems. SUS sentences are syntactically correct but have no meaning or present a very low predictability. The measure is the percentage of words that the listener properly recognizes.

Objective methods have been proposed to measure the quality of synthetic speech (Chu and Peng 2001; Falk and Moller 2008; Möller et al. 2010; Valentini-Botinhao et al. 2011), but given their low level of performance, these have not yet replaced perceptual methods (Hinterleitner et al. 2011, 2014; Norrenbrock et al. 2015).

1.6 Outline

In this paper, we describe the tasks related to the building of Emilia corpus. The main purpose of this corpus was to construct a female voice for Aromo, the TTS for Argentine Spanish (Torres et al. 2012). The Spanish spoken in Argentina, particularly in Buenos Aires city, is considered significantly different to other variations of Spanish (Colantoni and Gurlekian 2004; Coloma 2018). Section 2 provides a full description of all steps followed to construct the corpus, including validation and evaluation processes. Finally, Section 3 presents the conclusions and future work.

2 Emilia corpus building

The main features of Emilia voice are: unlimited text, real-time response, and high-quality synthesized speech similar to commercial TTS systems. The specification of real-time response is defined as: processing time for any text must be less or equal to the duration of synthesized speech.

Aromo has the usual modules: text pre-processing, prediction of intonation contours and segmental duration, unit selection and concatenation. The modules for prosody prediction are automatically close-fitting for each voice. The weights of the unit selection algorithm must be fixed manually for each voice (Torres et al. 2012).

In this section, the steps followed for the construction of Emilia corpus will be described, as presented in the previous section. The task of evaluation was performed in-house, during the process of corpus building.

2.1 Corpus specification

Aromo system is open domain: it has no restrictions on input text. Therefore, Emilia must have a coverage that makes it possible to synthesize any text input. Foreign

words will be spoken with Spanish phones, leaving the TTS system to apply the changes to these words that it deems appropriate in a Hispanicization process (Badino et al. 2004). Frequency distributions of unit must be similar to those found in local newspapers. Aromo is a unit selection TTS system, where the synthesis unit is the diphone. At least ten realizations of each concatenation unit are required, in line with the recommendations of Bonafonte et al. (2005). With this number of instances per unit, we attempt to achieve prosodic coverage with different phonetic contexts, stress conditions, and positions within the words and sentences. Given the exigency of processing time, the required corpus size should be kept to a minimum, ensuring coverage and high-quality synthesized speech. Umbert et al. (2006) sets a reference value of ten hours for a corpus of Iberian Spanish.

A professional female speaker from Buenos Aires city is required. The speech mode is read, including declarative and interrogative sentences. Speech must be perceived natural and pleasant, without monotony or too much expressiveness.

The recordings must be performed in an acoustic chamber with a noise level equal to or less than 30 dB SPL and the reverberation time should be RT60 $\leq$ 30 s.

The corpus must be labeled at the following levels: phonetic by Argentine SAMPA (Speech Assessment Methods: Phonetic Alphabet) (Gurlekian et al. 2001a), diphones, syllables, words, phrases and sentences, and part-of-speech (POS). The labeling shall be manual or semi-automatic with manual correction. The label file format will be ESPS/Waves+ (Entropic 1993). In addition, contours of fundamental frequency (F0), energy, and 15 coefficients of mel cepstrum have to be estimated and saved in plain text files.

2.2 Text corpus creation

Based on Sect. 1.2, we set out to define the text corpus based on two criteria: a sentence set was specifically designed to ensure coverage at various levels: phonetic, prosodic, common words, among others; a second sentence set was picked randomly from a reference text corpus representative of the application domain. This action plan seeks to capture the benefits of both approaches. This was implemented in three stages and with partial validations, as described below.

2.2.1 First stage

In a previous work (Gurlekian et al. 2001b) we had proposed to create a minimum set of sentences—exactly 741—which covered 97% of Spanish syllables in stressed and unstressed conditions. Both syllable conditions appeared in all allowed positions within the word. The main purpose of his design was aimed to study the prosody of sentences spoken at Buenos Aires resembling all kind of intonations. Particularly we aimed to study stress and accent relations within each intonative group. Texts were designed to represent syntactically simple and complex declarative sentences which resulted to contain one to six intonative groups after their production.

Sentences contained most frequent and common Argentine Spanish words to provide an everyday information or news style but also a set of additional items, such as business information, stock market data, weather forecast, traffic information, numbers from 0 to 100, fractions, isolated words and the isolated Spanish alphabet (Rodríguez 2000).

The general strategy was as follows: we started with more than 8000 sentences and only 520 sentences remained after filtering according to fixed criteria: sentence length: up to 20 syllables (Black and Lenzo 2003); presence of most of the possible syllables; in both stress conditions; and positions within the word: initial, medial, and final. The selection of the syllables was a two-step process. First, all the words in the most widely used Spanish dictionary (RAE 1992), and automatically divided into syllables. Syllables with only one token (5%) were left aside. Then, the presence of the most frequent syllables (700) was ensured by using (Guirao and Jurado 1993) frequency count. The selected syllables were then included in all the phonotactically allowed positions (word-initial, medial or final) (McPherson 1975) and with all the stress conditions (Harris 1983). Extra 221 sentences were created by two linguistic experts, who were instructed to make sentences with words containing less frequent syllables.

2.2.2 Second stage

In the second stage, the goal was to achieve a minimum coverage of five realizations of each diphone used in Argentine Spanish, keeping up the natural frequency distribution of diphone. Interrogative sentences were also included. First, the identity and relative distribution of the diphones usually performed in the Spanish spoken in Buenos Aires were estimated. To estimate a reference distribution, 2,896,666 sentences were extracted from local newspapers and were automatically transcribed to the Argentine SAMPA alphabet. One point to keep in mind is that not all diphones found come from Spanish words, and we had to filter the words manually. Another problem is the low rate of occurrence of some diphones, based on the LNRE problem, which also must be analyzed manually. Finally, the distribution of the diphone occurrences was estimated.

Then, sentences of five to ten words in length without foreign words were selected from the reference text. This is a semi-automatic iterative process, since we cannot identify a priori the diphones coming from foreign words. The Real Academia Española dictionary was used as reference. The sentence length restriction was imposed to equalize with first stage. The diphone histogram of the resulting sentences was tuned with the reference distribution. This process also took into account the occurrence of at least five realizations of each diphone. It is a very laborious task given the low rate of occurrence of certain units. To carry out this, Real Academia Española dictionary was used for substituting words in preselected sentences.

Finally, thinking about the possible TTS system applications, sentences with politeness connotation were added. The resulting sub-corpus had 652 declarative sentences and 200 interrogative sentences. Since the theoretical transcription from

Table 1 Statistics from the text corpus

	First stage	Second stage	Third stage	Final
Sentences	741	852	625	2218
Phrases	1224	1164	3874	6262
Words	5281	8210	15,542	29,033
Different words	2835	3055	4783	8925
Syllables	11,453	16,799	31,763	60,015
Different syllables	1413	995	2135	2641
Diphones	28,342	39,969	79,442	147,753
Different diphones	484	507	664	664
Phones plus allophones	27,118	38,806	75,570	141,494
Different phones plus allophones	30	30	30	30
Duration	37′25″	51′34″	105′34″	194′33″

graphemes to phones does not always match the spoken ones, after the sentences were recorded and labeled, we had to check out the gathered coverage later (see Sect. 2.4.2).

2.2.3 Third stage

After the second stage, an initial validation of the corpus was proposed. The sentences were recorded, tagged, and the voice was constructed (Torres et al. 2012), as described below. Then, speech quality of this first-version of Emilia voice was evaluated perceptually based on three standards (Gurlekian et al. 2012): P.85; SUS; and MOS. A detailed description of these tests and their results are presented below in Sect. 2.5.

An exhaustive analysis of the results of the perceptual tests was performed. The real distribution of the synthesis units was also re-estimated and was compared with the reference distribution. The errors found were tied in to: absence or scarcity of diphones, usually associated with some foreign words or fusion of rare words; phrases with structures absent in the corpus, for example, sentences with many intonative phrases, or sentences with short intonation phrases.

A comparison with synthetic voices of commercial TTS systems was also made. Due to the lack of an Argentine Spanish voice available to compare during the experiments, we used a Mexican Spanish (Paulina by Loquendo®) and a neutral Latin American Spanish voice (Rosa by ATT®). The results showed that the quality of this first version of Emilia voice was similar to that achieved by commercial systems (Gurlekian et al. 2012).

From this analysis, 625 new sentences were manually grouped into three subgroups. First, 36 interrogative sentences were added, including frequently asked questions in all possible phonetic endings. Second, 195 declarative sentences were added, specially designed to contain long sentences with prosodic structures not

Table 2 Statistics of diphones for Lana and Emilia text corpus

	Lana text corpus	Emilia text corpus
Diphones	728 (91.2%)	664 (83.2%)
Diphones freq. upper 0.01%	389 (48.7%)	463 (58.0%)
Diphones freq. [0.01 0.002]%	82 (10.3%)	149 (18.7%)
Diphones freq. [0.002 0.0005]%	52 (6.5%)	51 (6.5%)
Diphones freq. lower 0.0005%	205 (25.7%)	0 (0.0%)
Diphones missing	70 (8.8%)	135 (16.9%)

present in the corpus until now. Third, 393 sentences were added in order to achieve a minimum coverage of ten instances of each diphone. At this stage, every possible diphone in Spanish, including rare phone combinations, for example those from words in other languages and mergers of them, were contained in the corpus. These diphones have very low, if not null, appearance in the reference text corpus.

Dictionaries and word combinations were used to achieve the missing phone sequences, through a process of replacing words in sentences already in the corpus. Foreign words and proper names derived from other languages result in combinations of phones that do not normally occur in Spanish.

2.2.4 Final version

The statistics of the text sub-corpus generated in the three stages and the final version are shown in Table 1. This table shows how the number of different diphones increases in each stage: coverage was not taken into account in the first stage, and rare diphones were included in last stage. Both also affect the number of different words and syllables. The number of possible combinations of phones is 960, but they are not achieved in the corpus because: 163 diphones are not allowed, for example [nb] or [sl], and there were no instances of 133 diphones, such as [ww] or [bp]. The corpus has a diphone coverage of 83%.

The increase in the number of different syllables in the third stage corresponds to foreign words, mainly proper names inserted to achieve the required diphone coverage.

The total length of the recorded speech corpus is only three hours 14 minutes and 33 seconds, much less than other approaches (Umbert et al. 2006).

2.2.5 Diphone coverage validation

After the construction of the text corpus, a validation process of the synthesis unit coverage was performed to determine if the frequency of phones, allophones and diphones corresponded to the natural language. A new reference text corpus called Lana was built from a Buenos Aires newspaper collected during a ten years interval. This corpus was automatically transcribed into phones and allophones, using the

Argentine SAMPA code. This corpus consists of 55,908,107 words, of which 288,779 are different, with 270,214,192 phonemes.

The text corpora statistics of diphones are shown in Table 2, which confirms their LNRE distribution. Argentine SAMPA has 30 phones plus allophones and 960 possible phone combinations, but only 798 are allowed. In the Lana corpus, only 728 diphones were found, with 8.8% missing. The 99.99% events are covered with only 48.7% different diphones. Furthermore, 25.7% diphones have a frequency of occurrence lower than 0.0005%, 6.5% diphones have a frequency of occurrence between 0.0005% and 0.002% and 10.3% diphones fall in a range from 0.002% to 0.01%.

Table 2 also includes the statistics for the Emilia text corpus. The requirement of a minimum coverage of ten instances per diphone and all syllables in every stress position are the main causes of differences in distributions. Diphones with frequency of occurrence lower than 0.0005% were pushed up or down in distribution. This diphone set is coming from unusual foreign words. Because the TTS system must be able to pronounce any sequence of phones, for more unusual or strange sounds, a set of rules was created that makes it possible to supersede those units that do not meet the coverage requirements. These rules merge two consecutive equal phones, for example [tt] → [t], and separate them by inserting pauses into the two phones sequence difficult to pronounce in Spanish, e.g., [Jg] → [J g] (Torres 2012).

A Kolmogorov–Smirnov test was performed to determine whether normalized frequencies of phones and diphones in the corpora Lana and Emilia come from the same distribution, achieving a positive result with 99% of significance.

2.3 Speaker selection and recording

A female speaker was chosen to record the entire corpus. First, six professional female announcers, all natives of Buenos Aires, were evaluated through reading and recording ten paragraphs. The sentences presented the realizations of all phonemes and allophones. Both short and long intonative groups were designed to check intensity declinations and pitch variations. Then the selection was made by a group of interdisciplinary specialists: a phonetician who looked at quality production of each phone, an audiologist who listened to assess word intelligibility and a broadcaster to verify for pleasantness and fluency of their emissions.

The selected announcer was instructed to read the texts naturally as she did during the trial. Before recording a sentence, she had to read it in order to understand its meaning and plan her speech. The recording was made in an acoustic chamber with a noise insulation level of 30 dB SPL and a reverberation time of 0.2 s. An interactive program that displayed the text with further verification and/or playback of the recording was used. The recordings were made with a dynamic AKG microphone, sampling frequency of 16 kHz, 16-bit encoding and several short sessions to preserve speech quality.

2.3.1 Speech files validation

Each speech file was checked to ensure quality. First, a visual inspection of the speech wave was carried out to detect spikes, clippings and other recording noises. It was followed by a listening test to detect unnatural speech. Those speech files that did not pass this test were recorded again.

2.4 Corpus labelling

The voice of a TTS system by selection of units is composed of different modules, for example, units of synthesis and predictors of prosodic features. The modules of the synthesis units, in addition to the synthesis units themselves, have information of each unit, which is used by the process of unit selection. The module of prediction of prosodic features allows predicting from the text the contours of intonation, insertion of pauses and durations of each phone. Aromo uses artificial neural networks as predictors, which are trained for each voice with information extracted from Emilia corpus.

The speech files were labelled in five tiers: phones; diphones; syllables; words; and parts of speech.

A semi-automatic labelling procedure was accomplished: automatic labeling was followed by a manual correction by musically trained speech therapists (Gurlekian et al. 2014). A set of tools was implemented to perform the following tasks: grapheme to phone conversion (Gurlekian et al. 2001a); time alignment of phones, diphones, syllables, words and sentences; and part-of-speech labelling.

The phonetic automatic alignment was performed using the Hidden Markov Model Toolkit[1] (HTK). Since the monophones models showed better performances than triphones models on a reference subset of data, the alignment was performed using a set of 30 acoustic models based on the phones of the Argentine SAMPA, plus a unit for silence and another for short pauses. Each monophone model had three left to right emitting states, each with four Gaussian mixture components. The models were trained on the corpus to align. Lexical prompts were automatically transcribed to phones and were used to train the models. Phone networks were built and aligned at word and phone level using a forced alignment Viterbi search. The features set of the recognition system was obtained using 20 ms windows and five ms frame rate, and consisted of 12 Mel cepstral coefficients plus energy, and their delta and acceleration estimations.

The approach of Febrer et al. (1998) was used to align diphones. It used speech recognition based on demiphones (Marino et al. 2000). HTK was used with the same set of monophones, but now with each phone model containing four emitting states, and each state with four Gaussian mixture components with diagonal covariance matrices. Given the output of forced alignment, the diphones were obtained considering the segment from the transition between states two and three of a monophone to the transition from states two and three of the following phone.

[1] http://htk.eng.cam.ac.uk.

Syllable tiers were created automatically using the Real Academia Española rule set. The final versions of the lexical and phonetic transcriptions were used for this task.

Semi-automatic POS tagging was done with the Aromos tagger (Torres 2012), followed by manual correction performed by linguists. A reduced EAGLE[2] tag set[3] was used. EAGLE uses different attributes for each morphosyntactic category coded progressively. Only the first two levels of description were used: the morphosyntactic category and the first attribute. This level of detail is sufficient according to previous work (Torres and Gurlekian 2004, 2009).

F0, energy and mel cepstrum coefficients contour files were extracted with ESPS[4] get_f0 routine, Praat[5] software and Matlab®[6] routines, respectively. Parameters of Fujisakis intonation model (Fujisaki and Hirose 1984) were estimated using the Torres and Gurlekian (2016) algorithm.

2.4.1 *Labels file validation*

Each label file was checked to detect any irregularities or discrepancies in the labeling process. A tool (CTHTool) to arrange different labeling tiers as a hierarchical structure was built. The CTHTool verifies the presence of all the files, the consistency of the labeling of each level, and if the labels are allowed. It also seeks discrepancies between automatic grapheme to phone transcription and manual correction. A report indicating possible errors is generated as output, and then a manual check on the label files must be performed. This tool also extracts all information needed to build the TTS voice from the corpus.

The F0 contour files were also analyzed. The F0 estimation method is robust under the conditions in which the recordings were made, but some errors may occur in the event of a weak energy signal in the micro-prosody and doubling/halving effects. The Fujisaki model was used to detect possible errors in F0 contour estimated. For each F0 contour file, Fujisakis parameters model was extracted using Torres and Gurlekian (2016) method. Then, the F0 contours by model were generated, and they were compared with the extracted model from the speech wave. Finally, files with the biggest differences were analyzed manually, in terms of average and absolute values point to point.

2.4.2 *Final coverage validation*

The theoretical coverage based on an automatic transcription of grapheme to phone was validated in Sect. 2.2.5. But there are differences between the phones that the speaker should emit in theory and those that are emitted in the end. This occurs for several reasons: insertion of pauses not marked in the text, bad habits of the speaker,

[2] http://www.ilc.cnr.it/EAGLES96/home.html.

[3] http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html.

[4] http://www.speech.kth.se/software/.

[5] http://www.fon.hum.uva.nl/praat/.

[6] http://www.mathworks.com/.

alternative pronunciations, among others. Therefore, it is necessary to make a final assessment of the corpus coverage based on the corrected transcript of the speech files.

The number of occurrences of each phone, according to SAMPA alphabet, for Lana, Emilia text and speech corpora are shown in Fig. 1. The phone with the highest number of appearances is [e], with 14,029 instances, and [g] is the one with less occurrences with only 159 cases. Open vowels are more frequent than closed ones. Phonemes /s/, /r/ and /n/ appear with high occurrences due to their final position in the word. In addition, /s/ creates plurals and /l/ forms most of the determinants. This graph shows the disparity degree in the number of occurrence of each phone in the corpus. This reflects previous works conducted by Guirao and Jurado (1993).

Figure 2 shows the number of diphones grouped by ranges of the number of occurrences in the Emilia speech corpus. Boundary values of each range were defined a priori in order to illustrate the distribution. The diphone group with occurrences less or equal to 15 are 139, which comprises the less probable diphones, which occur only in very specific situations, such as diphones generated by the fusion of nouns that end in a consonant and that follow a word that begins with a consonant, for example the diphone [fn] in the context of the words ... *Budasof nació*... (... *Budasof was born*...). Andersen and Hoequist (2003) have already reported this type of behavior in their work in Danish. The group of 37 diphones with more than 1000 occurrences each is composed of high frequency diphones in speech, such as [el] [es] [en] [la], among others.

In Table 3 the number of occurrences of each diphone is shown in percentages. Rows are the identity of the first phone and columns are the second phone. Almost three-quarters (73.06%) of possible diphones are well represented in Emilia according the reference corpus. Diphone distributions were analyzed in groups as shown in Table 2 . A font color code was used to identify each group in Table 3. The cells corresponding to missing diphones were filled with a color square. Red squares indicate non-realizable diphones in Spanish from Argentina.

- Diphones with occurrence frequency above 0.01% (fonts in black color in Table 3). Three diphones are missing: [DD], [tt] and [kk] (grey squares in Table 3). Double consonant phones are merged by speaker. As a way to maintain coverage, new phone merge rules were added to the TTS system.
- Diphones with an occurrence frequency between [0.01 0.002] %, in blue color in Table 3. Fourteen diphones are missing. Eight are double phones that merge as in the previous case. Six are consonant combinations, between which the speaker inserts a short pause. As a way to maintain coverage, new phone rules were added to the TTS system.
- Diphones with an occurrence frequency between [0.002 0.0005] %, in brown color in Table 3. Nine diphones are missing. Idem previous item: one double phone merged; and eight rare consonant combinations. These were found in the context of foreign words or a combination of them. To ensure coverage, phone transformation rules were added to the TTS system, inserting short pauses or vowels based on phonetic criteria.

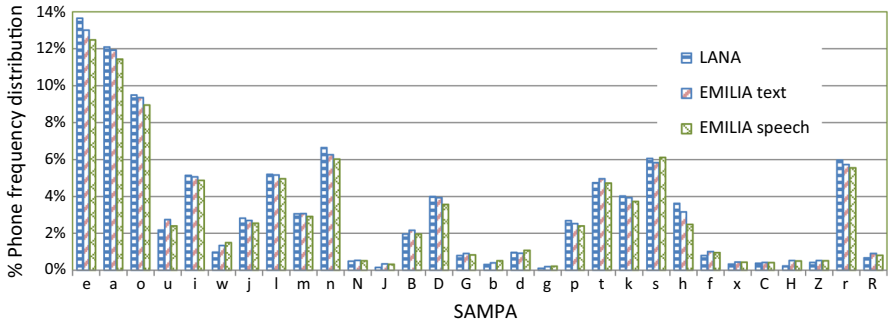


Fig. 1 Phoneme frequency distribution in Lana, Emilia text and speech corpora, in %

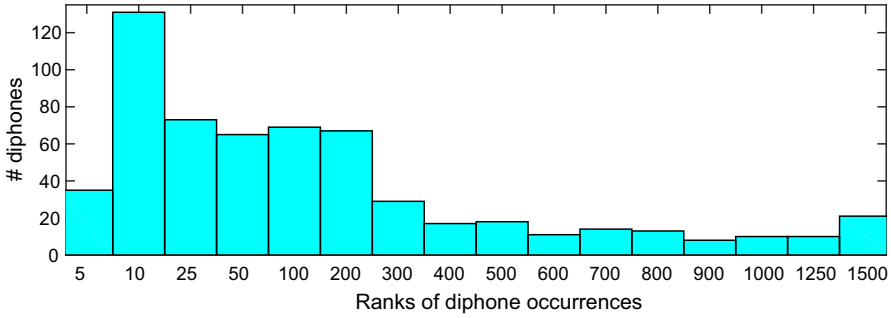


Fig. 2 Number of diphones grouped by ranges of occurrences number in Emilia speech corpus

- Diphones with an occurrence frequency less than 0.0005%, in cyan color in Table 3. It is the largest group of phones missing: 129. They correspond to foreign words or a combination of them. To ensure coverage, phone transformation rules were added to the TTS system, inserting short pauses or vowels based on phonetic criteria.
- Missing diphones. They are a group of remaining diphones that were not found in the reference corpus or Emilia: 60; in green color in Table 3. These correspond to sequences of consonants that do not exist in Spanish. To ensure coverage, phone transformation rules were added to the TTS system, inserting short pauses or vowels based on phonetic criteria.

2.5 Corpus evaluation

2.5.1 Time response measurement

Preliminary tests show that the conversion speed depends on the length of the text to convert. The Aromo TTS system is responsible for these variations (Torres 2013; Torres and Gurlekian 2016).

Table 3 Distribution of the occurrences of the diphones in the data corpus, in %

Identity of the first phone	Identity of the second phone																														
	e	a	o	u	i	w	j	l	m	n	N	J	B	D	G	b	d	g	p	t	k	s	h	f	x	C	H	Z	r	R	–
e	0.313	0.440	0.131	0.021	0.032	0.137	0.241	1.580	0.508	1.936	0.129	0.089	0.396	0.563	0.231				0.263	0.421	0.648	1.360	0.906	0.135	0.074	0.092	0.060	0.056	1.230	0.160	0.414
a	0.322	0.216	0.090	0.020	0.124	0.229	0.197	1.095	0.556	0.948	0.076	0.075	0.510	1.157	0.125				0.381	0.448	0.606	1.225	0.468	0.150	0.132	0.080	0.038	0.137	1.232	0.191	0.668
o	0.304	0.220	0.072	0.032	0.057	0.112	0.092	0.417	0.585	1.260	0.092	0.025	0.318	0.504	0.085				0.307	0.226	0.453	1.087	0.568	0.103	0.045	0.025	0.144	0.042	0.890	0.091	0.833
u	0.027	0.031	0.010	0.005	0.009	0.007	0.037	0.181	0.160	0.523	0.034	0.012	0.170	0.128	0.037				0.101	0.091	0.103	0.154	0.125	0.028	0.036	0.015	0.058	0.040	0.210	0.037	0.029
i	0.054	0.264	0.039	0.014		0.011	0.005	0.336	0.273	0.535	0.136	0.016	0.262	0.464	0.113				0.107	0.346	0.451	0.544	0.293	0.102	0.021	0.036	0.042	0.079	0.255	0.023	0.065
w	0.667	0.277	0.025		0.057		0.014	0.011	0.043	0.180	0.027		0.008	0.024	0.008				0.008	0.042	0.013	0.021	0.021	0.009		0.005	0.010	0.007	0.027		0.012
j	0.698	0.473	0.813	0.012		0.025		0.018	0.064	0.178	0.023		0.011	0.046	0.014				0.009	0.018	0.027	0.043	0.019	0.007	0.005	0.006	0.006	0.005	0.016	0.005	0.033
l	0.622	1.367	0.624	0.148	0.435	0.045	0.075	0.031	0.123	0.051		0.007	0.105		0.068		0.023		0.148	0.216	0.196	0.140		0.066	0.015	0.014	0.014	0.016		0.062	0.239
m	0.501	0.581	0.396	0.150	0.355	0.014	0.082	0.005	0.007	0.013		0.007				0.253	0.008	0.006	0.346	0.005	0.010			0.123			0.007	0.004		0.005	0.045
n	0.674	0.772	0.689	0.160	0.311	0.172	0.062	0.211	0.085	0.053		0.007					0.531			1.266		0.568	0.018			0.029	0.036	0.013		0.055	0.322
N																		0.157			0.325				0.020						
J	0.028	0.112	0.142	0.020	0.020	0.007																									
B	0.447	0.396	0.214	0.058	0.268	0.029	0.119	0.144		0.005				0.011	0.005				0.010	0.011	0.010	0.013	0.012	0.012	0.005	0.010	0.008	0.009	0.156		0.029
D	1.382	0.588	0.744	0.060	0.337	0.018	0.137	0.009	0.017	0.012		0.006	0.012		0.008				0.014	0.012	0.014	0.014	0.005	0.014	0.010	0.010	0.012	0.006	0.053	0.012	0.089
G	0.025	0.205	0.171	0.075	0.031	0.047	0.012	0.022	0.012	0.021		0.007	0.004	0.005						0.006	0.005	0.007		0.007			0.007		0.140		0.023
b	0.084	0.066	0.043	0.027	0.052	0.060	0.070	0.020												0.004		0.010	0.004						0.064		
d	0.488	0.099	0.218	0.034	0.110	0.015	0.055																						0.018		0.003
g	0.008	0.051	0.036	0.026	0.007	0.014	0.005	0.015														0.004							0.044		0.010
p	0.374	0.531	0.516	0.127	0.101	0.094	0.054	0.129		0.005		0.006			0.004					0.036	0.007	0.023		0.005			0.007	0.003	0.375		0.027
t	0.982	1.129	0.873	0.199	0.527	0.039	0.163	0.013	0.011	0.009		0.005	0.011	0.008	0.006				0.007		0.010	0.015		0.011	0.006		0.010	0.003	0.681		0.049
k	0.583	0.698	1.163	0.174	0.129	0.229	0.056	0.091	0.006	0.015		0.006		0.008					0.008	0.226		0.153	0.042	0.007		0.006	0.007		0.122	0.006	0.029
s	1.329	0.688	0.558	0.602	0.814	0.036	0.824																								0.978
h								0.092	0.150	0.088		0.006	0.058	0.257	0.047				0.348	0.970	0.354	0.007	0.000	0.024	0.008	0.007		0.007		0.042	
f	0.131	0.104	0.129	0.052	0.204	0.103	0.027	0.047		0.005		0.007		0.005						0.009	0.010				0.006			0.006	0.099		0.017
x		0.105	0.118	0.076		0.020		0.008	0.008	0.007		0.007	0.008	0.007	0.006					0.007	0.007				0.008	0.008	0.006	0.005		0.007	0.018
C	0.276				0.105		0.027	0.005																							0.007
H	0.088	0.138	0.130	0.019	0.052	0.008	0.007					0.005								0.005				0.005		0.005					0.004	0.040
Z	0.118	0.151	0.092	0.026	0.023	0.007	0.004	0.006	0.008	0.005		0.006	0.005		0.004				0.007	0.006	0.005	0.005					0.007				0.027
r	0.853	1.310	0.820	0.125	0.385	0.016	0.166	0.191	0.193	0.090		0.008	0.074	0.196	0.066				0.048	0.238	0.160	0.246	0.007	0.031	0.014	0.039	0.013	0.014		0.016	0.279
R	0.491	0.105	0.088	0.032	0.059	0.010	0.024	0.004		0.003																					
-	0.750	0.453	0.114	0.136	0.322	0.012	0.025	0.318	0.122	0.131		0.023				0.161	0.203	0.031	0.287	0.136	0.320	0.422	0.010	0.068	0.027	0.025	0.017	0.063		0.094	

See Sect. 2.4.2 for color references

From the Lana corpus, 80 sentences were randomly selected, divided into two groups based on the following criterion: the first group consisted of sentences of five to 15 words in length; and the second group contained sentences of 29 to 35 words in length. The duration of the synthesized audios ranged from four to seven seconds and 14 to 18 seconds, respectively.

Tests were performed on a desktop computer, with an Intel (R) Core i7-2600k CPU @ 3.40 GHz microprocessor and 8 GB of RAM, with a Windows 7(R) operating system, Service Pack 1.

The xRT scale was estimated: ratio between the processing times and the durations of the synthesized speech. The short sentences xRT oscillate between one and 0.8. For long sentences, the xRT were in the range of 0.5 to 0.6.

2.5.2 Synthesized speech quality

The corpus was assessed by conducting a perceptual evaluation of the synthesized speech. Two tests were carry out: in the first one (1st Test) the voice was built with the data of the first and second stage of corpus (see Sect. 2.2); and in the second one (2nd Test) the entire corpus was used. Four perceptual tests were conducted: MOS, ITU P85, SUS and CMOS between the voices of the 1st Test and 2nd Test.

Speech Language Therapists (Experts) and non-expert (Naïve) subjects without auditory impairments participated in the experiment. In the first test, ten experts and ten naïve listeners participated. In the second test, 12 experts and 18 naïve listeners participated. The same text sentences were used in the first and second tests. Thirteen experts and 13 naïve listeners were used in CMOS test.

ITU Test Nine different texts were created, 20–25 words in length with up to ten short melodic phrases. Texts were designed for three tasks: telephone sales, flight information and payment services. The sentences have a fixed portion specific to the task and a variable part that changes in each presentation producing several melodic groups. Foreign proper names were included. See Gurlekian et al. (2012) for the complete text of the stimuli.

Listeners could hear stimuli as often as required to make judgments of intelligibility (i) and quality (q). Below is a summary of the instructions given to the subjects, the scales used and in which part they appeared.

- Word Recognition (i): *Please write down [name, product, features, code, price and time], [company, country, flight number, departure time, terminal and boarding gate] or [name, company, month, amount and branch] (according to each task).*
- Overall Impression (q): *How would you rate the quality of what you heard? Excellent; Good; Fair; Poor; Bad.*
- Listening Effort (i): *How would you describe the effort needed to understand the message? None; Low; Moderate; High; Not understood with any amount of effort.*
- Comprehension Problems (i): *Were there words that were difficult to understand? Never; Rarely; Occasionally; Often; All the time.*

- Articulation (i): *Were the sounds distinguishable? Yes, clear; Yes, clear enough; Fairly clear; Not very clear; Not at all.*
- Pronunciation (q): *Did you note any anomalies in the pronunciation? No; Yes, but they were not annoying; Yes, slightly annoying; Yes, annoying; Yes, very annoying.*
- Speaking Rate (q): *The average speed was: Much faster than preferred; Faster than preferred; Preferred; Slower than preferred; Much slower than preferred.*
- Voice Pleasantness: (q) *How would you describe the voice? Very pleasant; Pleasant; Fair; Unpleasant; Very unpleasant.*
- Acceptance (i, q): *Do you think that this voice could be used for an information service by telephone? Yes; No.*

SUS Test Fifty semantically nonsensical texts consisting of six to ten words were designed using correct syntactic structure; each contained one or two melodic phrases. For each listener, 15 randomized sentences without repetitions were employed. Instruction given subjects was: *Write down each word that you hear in the sentence. The recording will not be replayed, so be alert*. See Gurlekian et al. (2012) for the complete text of the stimuli.

MOS Test Sixty texts, each consisting of ten to 20 words in two or three melodic phrases in length, were synthesized. Listeners evaluated the quality of synthesized speech on a ten-point scale. Subjects listened to five sentences without repetitions. See Gurlekian et al. (2012) for the complete text of the stimuli.

CMOS Test MOS test stimuli from the first and second tests were used in pairs. These were randomly chosen and presented to the listeners at random. Each subject evaluated 15 randomly selected stimuli. Instruction given subjects was: *Compare the second stimulus against the first, and select an option to rate it*. The scale used to rate stimuli was: 3 (Much better); 2 (Better); 1 (Slightly better); 0 (About the same); −1 (Slightly worse); −2 (Worse); −3 (Much worse).

Results Two Emilia voice were built and speech stimuli were synthesized with Aromo TTS system for the first and second tests. The tests were carried out under the supervision of specialized personnel, using a platform developed for this task. The test results are shown in Table 4 . Standard deviation has been included as a measure of dispersion. Test results of SUS, ITU Word Recognition and ITU Acceptance are expressed in percentage. The ITU standard indicates that a variance analysis should be performed. ANOVA test could not be performed since the data did not pass Lillie's normality test, nor did the majority pass Levene's homoscedasticity test. Therefore, the Kruskal–Wallis nonparametric test was performed instead; its results are also shown in Table 4. The boxplots are shown in Fig. 3.

The results of all tests show a high intelligibility and naturalness of synthesized speech, both in the first and second tests conducted, and among naïve and expert listeners alike. When comparing the first and second test results for naïve and expert listeners, we found some significant differences. The MOS for naïve listeners improved by 28% when comparing the results of the first and second tests. This improvement is also reflected in the results of the ITU Voice Pleasantness/ Acceptance and CMOS test. This upturn can be attributed to the improvement in

Table 4 Results of perceptual tests

	1st test		2st test		Kruskal–Wallis test			
	Expert	Naïve	Expert	Naïve	Expert		Naïve	
					χ^2	$p > \chi^2$	χ^2	$p > \chi^2$
MOS	7.06 ± 1.41	6.22 ± 1.82	7.18 ± 2.10	7.94 ± 1.77	0.4	0.600	25.6	0.000
SUS %	95 ± 9	94 ± 12	99 ± 4	98 ± 6	21.7	0.000	12.7	0.000
ITU word recognition %	91 ± 16	86 ± 15	89 ± 14	89 ±12	2.2	0.138	0.03	0.870
ITU overall impression	3.13 ± 0.94	3.23 ± 0.90	3.78 ± 0.83	3.52 ± 0.77	8.1	0.004	2.3	0.128
ITU listening effort	3.90 ± 0.90	3.93 ± 0.90	3.97 ± 0.92	3.54 ± 0.88	0.1	0.733	3.7	0.054
ITU pronunciation	3.63 ± 1.03	3.87 ± 0.78	4.08 ± 0.94	3.67 ± 0.80	3.2	0.072	1.1	0.294
ITU comprehension problems	4.14 ± 0.83	3.82 ± 0.72	4.03 ± 0.84	3.93 ± 0.80	0.3	0.574	0.3	0.591
ITU speaking rate	4.27 ± 1.11	4.40 ± 0.93	4.67 ± 0.76	4.55 ± 0.85	2.6	0.107	0.5	0.462
ITU articulation	4.00 ± 0.71	4.00 ± 0.90	3.53 ± 1.08	3.72 ± 0.98	3.4	0.063	1.8	0.180
ITU voice pleasantness	3.63 ± 0.76	3.27 ± 0.94	4.00 ± 0.89	3.78 ± 0.90	3.5	0.062	5.4	0.021
ITU acceptance %	63 ± 49	60 ± 49	86 ± 35	80 ± 41	6.0	0.015	4.9	0.27
CMOS			0.74 ± 1.38	0.68 ± 1.42				

Standard deviation has been included as a measure of dispersion

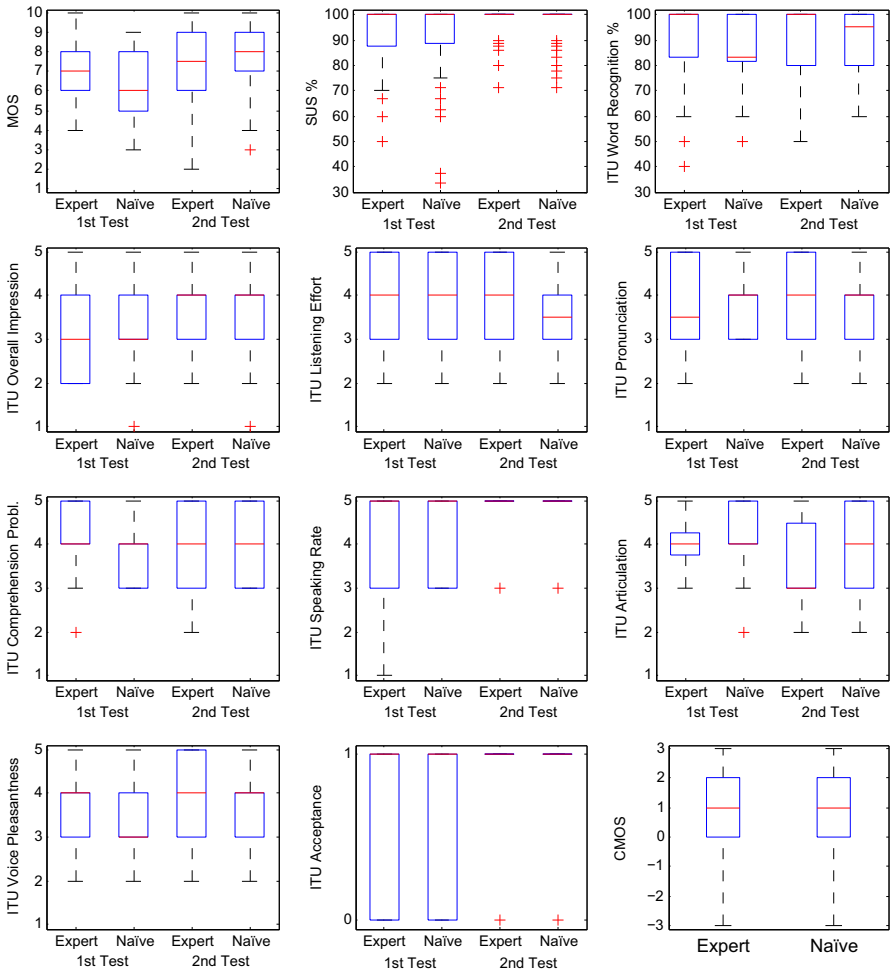


Fig. 3 Boxplot of perceptual evaluation results: MOS; SUS; ITUs; and CMOS

synthesized speech prosody in the prediction stage of prosodic features and in the realization of the prosodic contours. The results of the SUS test are conclusive: in the first test, values close to natural speech are achieved, and in the second test, values close to those obtained in natural-speaking sentences with meaning were improved (Aguilar et al. 1994; Sainz et al. 2010). The results obtained for the ITU test are similar to those obtained for commercial TTS systems in other variants of Spanish and other languages (Fernández-Torné and Matamala 2015; Alvarez and Huckvale 2002).

3 Conclusions and future works

This paper describes the design and construction of a speech corpus for building a voice for a TTS system for Spanish spoken in Buenos Aires. The goal was to obtain a high-quality voice with the smallest data corpus possible. The strategy for the design of text corpus was oriented to start with a small data set that fulfilled part of the specification, and then, in successive stages to increase the number of sentences in order to meet any additional requirements. An evaluation between the stages made it possible to detect and correct corpus failures. The speech corpus obtained meets all of the requirements originally stated. Its size is relatively small, half of the recommended size for these applications (Bonafonte et al. 2006; Taylor 2009), and the results of the perceptual evaluation test indicate a high quality of synthesized speech.

The text of the corpus was created by language masters, with the support of dictionaries and texts extracted from local newspapers. Part of the text was automatically extracted from the reference corpus, but another had to be created or adapted manually to ensure the required coverage of the units, lexical stress conditions and prosodic structures, as well as minimize their size. The LNRE distribution makes its presence strong, making the task more difficult.

A set of resources and tools were developed or adapted to carry out this work: a large corpus of reference text; a text corpus to perform perceptual tests; text selection tools; forced phoneme alignment tool; syllabification tool; verification tools of labeling consistency; tools for the extraction of fundamental frequency and energy; tool for prosodic labeling; and tool for labeling part-of-speech.[7]

The task of labeling requires great effort. In this task we chose to perform a semiautomatic labeling: a first automatic approach with a subsequent manual correction stage. Even when the manual correction is performed by highly trained and experienced labelers, which is an expensive and time-consuming task, the results obtained must be analyzed to ensure consistency.

The quality of synthesized speech depends on both the corpus and the TTS system. Many of the problems associated with selecting units would be solved if we knew how to measure the quality of a sentence, first perceptually and then by implementing automatic methods. Moreover, it is still not clear how and what should be asked of perceptual quality tests. In recent works, the validity of the results obtained in perceptual tests of the synthesized speech quality has been questioned harshly. Betz et al. (2018) mentioned that the quality perceived by the users depends on the application where the TTS system is implemented, and they propose to evaluate the speech quality generated by the TTS systems in the context of each application. The work of Mendelson and Aylett (2017) also supports this conception. This approach has several disadvantages: it is expensive, slow and the results will depend on the profile of the respondents who perform the test (Cryer and Home 2010). Rosenberg and Ramabhadran (2017) show that the MOS tests results

[7] These resources are available, in full, partial or demonstrations, for academic or commercial purpose(s), by e-mail to the authors.

are subject to bias from different sources, and this bias must be eliminated to obtain more representative results.

The distribution of diphones occurrences shown in Table 3 is as expected, but it is fascinating that there is still no theory of speech production that can explain it.

Much effort was made in the design of the sentences, but they have the advantage of being reusable for the construction of other voices. A new male voice is currently under construction using the Emilia text corpus.

Acknowledgements The authors would like to thank the anonymous reviewers for their insightful feedback. This research was supported by Ministerio de Ciencia y Tecnología and Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina.

References

Adell, J., Bonafonte, A., Gomez J., & Castro, M. (2005). Comparative study of automatic phone segmentation methods for TTS. In *Proceedings of the ICASSP'05*, (pp. 309–312). https://doi.org/10.1109/ICASSP.2005.1415112.

Aguilar, L., Fernzández, J., Garrido J., Llisterri, J., Monzón, A. M. L., & Crespo, M. R. (1994). Evaluation of a Spanish text-to-speech system. In *Proceedings of the second ESCA/IEEE workshop on speech synthesis* (pp. 207–210). https://www.isca-speech.org/archive_open/archive_papers/ssw2/ssw2_207.pdf.

Alıas, F., Iriondo, I., & Barnola, P. (2003). Multi-domain text classification for unit selection text-to-speech synthesis. In *Procedings of the 15th international congress of phonetic sciences* (pp. 2341–2344). https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/papers/p15_2341.pdf.

Alvarez, Y. V., & Huckvale, M. (2002). The reliability of the ITU-T P.85 standard for the evaluation of text-to-speech systems. In *Proceedings of the 7th international conference on speech & language processing* (pp. 329–332). https://www.isca-speech.org/archive/archive_papers/icslp_2002/i02_0329.pdf.

Andersen, O., & Hoequist, C. (2003). Keeping rare events rare. In *Proceedings of the eighth European conference on speech communication & technology* (pp. II-1337–II-1340). https://www.isca-speech.org/archive/archive_papers/eurospeech_2003/e03_1337.pdf.

Badino, L., Barolo, C., & Quazza, S. (2004). Language independent phoneme mapping for foreign TTS. *Proceedings of the fifth ISCA workshop on speech synthesis, Pittsburgh, PA, USA* (pp. 127–137). https://www.isca-speech.org/archive_open/archive_papers/ssw5/ssw5_217.pdf.

Bayerl, P. S., & Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, *37*(4), 699–725. https://doi.org/10.1162/COLI_a_00074.

Bellegarda, J. R. (2008). Unit-centric feature mapping for inventory pruning in unit selection text-to-speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(1), 74–82. https://doi.org/10.1109/TASL.2007.911059.

Benoît, C., Grice, M., & Hazan, V. (1966). The SUS test: A method for the assessment of TTS synthesis intelligibility. *Speech Communication*, *18*(4), 381–392. https://doi.org/10.1016/0167-6393(96)00026-X.

Betz, S., Carlmeyer, B., Wagner, P., & Wrede, B. (2018). Interactive hesitation synthesis: Modelling and evaluation. *Multimodal Technologies and Interaction*, *2*(1), 9. https://doi.org/10.3390/mti2010009.

Beutnagel, M., & Conkie, A. (1999). Interaction of units in a unit selection database. In *Proceedings of the sixth European conference on speech communication and technology* (Vol. 3, pp. 1063–1066). https://www.isca-speech.org/archive/archive_papers/eurospeech_1999/e99_1063.pdf.

Black, A. W., & Lenzo, K. A. (2000). Limited domain synthesis. *Proceedings of the 6th international conference on spoken language processing* (Vol. 2, pp. 411–414). https://www.isca-speech.org/archive/archive_papers/icslp_2000/i00_2411.pdf.

Black, A. W., & Lenzo, K. A. (2003). Building synthetic voices. Language Technologies Institute, Carnegie Mellon University and Cepstral LLC 4:2. http://festvox.org/bsv/bsv.pdf.

Boëffard, O. (2001). Variable-length acoustic units inference for text-to-speech synthesis. In *Proceedings of the 7th European conference on speech communication and technology* (pp. 983–986). https://www.isca-speech.org/archive/archive_papers/eurospeech_2001/e01_0983.pdf.

Bonafonte, A., Höge, H., Kiss I., Moreno, A., Ziegenhain, U., Heuvel, H., Hain, H., Wang, X., & Garcia, M. (2006). TC-STAR: Specifications of language resources and evaluation for speech. In *Proceedings of the 5th interantional conference on language resources and evaluation* (pp. 311–314). http://nlp.lsi.upc.edu/publications/papers/tc_star_spec.pdf.

Bonafonte, A., Höge, H., Tropf, H. S., Moreno, A., van der Heuvel, H., Sündermann, D., Ziegenhain, U., Kiss, J. P. I., & Jokisch, O. (2005). TTS baselines and specifications. In *Deliverable D8 of the EU project TC-STAR technology and corpora for speech to speech translation (FP6-506738)*. http://nlp.lsi.upc.edu/publications/papers/tc_star_spec.pdf.

Bozkurt, B., Ozturk, O., & Dutoit, T. (2003). Text design for TTS speech corpus building using a modified greedy selection. In *Proceedings of the eighth European conference on speech communication and technology* (pp. 277–280). https://www.isca-speech.org/archive/archive_papers/eurospeech_2003/e03_0277.pdf.

Breen, A. P., & Jackson, P. (1998). Non-uniform unit selection and the similarity metric within BT's laureate TTS system. In *Proceedings of the third ESCA workshop on speech synthesis* (pp. 373–376). https://www.isca-speech.org/archive_open/archive_papers/ssw3/ssw3_201.pdf.

Campbell, N. (1996). Chatr: A high-definition speech re-sequencing system. In *Proceedings of the 3rd ASA/ASJ joint meeting* (pp. 1223–1228). http://www.speech-data.jp/nick/feast/proceeding/asa-asj%201996_12.pdf

Campbell, N. (2005). Developments in corpus-based speech synthesis: Approaching natural conversational speech. *IEICE Transactions on Information and Systems*, *88*(3), 376–383. https://doi.org/10.1093/ietisy/e88-d.3.376.

Chalamandaris, A., Tsiakoulis, P., Raptis, S., & Karabetsos, S. (2011). Corpus design for a unit selection TTS system with application to Bulgarian. *Human Language Technology Challenges for Computer Science and Linguistics*, *6562*, 35–46. https://doi.org/10.1007/978-3-642-20095-3_4.

Chevelu, J., Barbot, N., Boeffard, O., & Delhay, A. (2008). Comparing set-covering strategies for optimal corpus design. In *Proceedings of the 23rd European signal processing conference* (pp. 2951–2956). http://lrec-conf.org/proceedings/lrec2008/pdf/750_paper.pdf.

Chevelu, J., & Lolive, D. (2015). Do not build your TTS training corpus randomly. In *Proceedings of the signal processing conference, IEEE* (pp. 350–354). https://doi.org/10.1109/EUSIPCO.2015.7362403.

Chu, M., Chen, Y., Zhao, Y., Li, Y., & Soong, F. (2006). A study on how human annotations benefit the TTS voice. In *Proceedings of the blizzard challenge workshop 2006*. http://www.festvox.org/blizzard/bc2006/msra_blizzard2006.pdf.

Chu, M., & Peng, H. (2001). An objective measure for estimating MOS of synthesized speech. In *Proceedings of the eventh European conference on speech communication and technology* (Vol. 3, pp. 2087–2090). https://www.isca-speech.org/archive/archive_papers/eurospeech_2001/e01_2087.pdf.

Coelho, L., Hain, HU., Jokisch, O., & Braga, D. (2009). Towards an objective voice preference definition for the portuguese language. In *Proceedings of the joint SIG-IL/microsoft workshop on speech and language technologies for Iberian languages* (pp. 67–70). http://www.isca-speech.org/archive_open/sltech_2009/papers/isl9_067.pdf.

Colantoni, L., & Gurlekian, J. (2004). Convergence and intonation: Historical evidence from Buenos Aires Spanish. *Bilingualism: Language and Cognition*, *7*(2), 107–119. https://doi.org/10.1017/S1366728904001488.

Coloma, G. (2018). Illustrations of the IPA: Argentine Spanish. *Journal of the International Phonetic Association*, *48*, 243–250. https://doi.org/10.1017/S0025100317000275.

Cryer, H., & Home, S. (2010). Review of methods for evaluating synthetic speech. RNIB Centre for Accessible Information, Birmingham: Technical report #8. https://www.rnib.org.uk/sites/default/files/2010_02_Evaluating_synthetic_speech_review.doc.

Dutoit, T. (1997). *An introduction to text-to-speech synthesis. Text, speech and language technology*. Dordrecht: Kluwer Academic.

Dybkjær, L., & Hemsen, H. (2007). *Evaluation of text and speech systems*. Berlin: Springer.

Eisen, B. (1993). Reliability of speech segmentation and labelling at different levels of transcription. In *Proccedings of 3rd European conference on speech communication and technology* (Vol. 1, pp. 673–676). https://www.isca-speech.org/archive/archive_papers/eurospeech_1993/e93_0673.pdf.

Entropic. (1993). *ESPS version 5.0 programs manual*. Washington, D.C.: Entropic Research Laboratory.

Falk, T. H., & Moller, S. (2008). Towards signal-based instrumental quality diagnosis for text-to-speech systems. *IEEE Signal Processing Letters*, *15*, 781–784. https://doi.org/10.1109/LSP.2008.2006709.

Febrer, A., Padrell, J., & Bonafonte, A. (1998). Generation of unit databases for the UPC text-to-speech system. In *Proceedings of the international workshop on speech and computer* (pp. 26–29). http://www.lsi.upc.edu/∼nlp/papers/febrer98b.pdf.

Fernández-Torné, A., & Matamala, A. (2015). Text-to-speech vs. human voiced audio descriptions: A reception study in films dubbed into catalan. *The Journal of Specialised Translation, 24*, 61–88. http://www.jostrans.org/issue24/art_fernandez.php.

François, H., & Boëffard, O. (2001). Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem. In *Proceedings of the seventh European conference on speech communication and technology* (pp. 829–832). https://www.isca-speech.org/archive/archive_papers/eurospeech_2001/e01_0829.pdf

François, H., & Boëffard, O. (2002). The greedy algorithm and its application to the construction of a continuous speech database. In *Procedings of the third international conference on language resources and evaluation* (pp. 1420–1426). http://lrec.elra.info/proceedings/lrec2002/pdf/265.pdf.

Fujisaki, H. & Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of japanese. *Journal of Acoustic Society of Japan, 5*(4), 233–242. https://www.jstage.jst.go.jp/article/ast1980/5/4/5_4_233/_pdf.

Grůber, M., Matoušek, J., Tihelka, D., & Hanzlicek, Z. (2014). Reducing footprint of unit selection TTS system by removing linguistic segments with rarely selected units. In *Proceedings of the 12th international conference on signal processing* (pp. 494–499). https://doi.org/10.1109/ICOSP.2014.7015054

Grůber, M., Tihelka, D., & Matoušek, J. (2007). Evaluation of various unit types in the unit selection approach for the czech language using the festival system. In *Proceedings of the 6th ISCA workshop on speech synthesis* (pp. 276–281). http://www.isca-speech.org/archive_open/archive_papers/ssw6/ssw6_276.pdf.

Guirao, M., & Jurado, M. G. (1993). *Estudio estadístico del español*. Buenos Aires: CONICET.

Gurlekian, J. A., Colantoni, L., & Torres, H. M. (2001a). El alfabeto fonético SAMPA y el diseño de córpora fonéticamente balanceados. *Fonoaudiológica*, *47*(3), 58–70.

Gurlekian, J. A., Cossio-Mercado, C., Torres, H. M., & Vaccari, M. E. (2012). Subjective evaluation of a high quality text-to-speech system for argentine spanish. In *Proceedings of VII Jornadas en Tecnologí del Habla and III Iberian SLTech Workshop, IberSPEECH 2012* (pp. 241–250). https://www.researchgate.net/profile/Christian_Cossio-Mercado/publication/265955190_Subjective_Evaluation_of_a_High_Quality_Text-to-Speech_System_for_Argentine_Spanish/links/552ef53d0cf2acd38cbbdad4.pdf.

Gurlekian, J. A., Rodríguez, H., Colantoni, L., & Torres, H. M. (2001b). Development of a prosodic database for an argentine spanish text to speech system. In B. Bird, & M. Liberman (Eds.) *Proceedings of the IRCS workshop on linguistic databases, SIAM* (pp. 99–104). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.25.5050&rep=rep1&type=pdf.

Gurlekian, J. A., Torres, H. M., & Evin, D. (2014). Guía para la segmentación y transcripción fonética para las tecnologías del habla. *Fonoaudiológica*, *61*(2), 24–27.

Hall, J. L. (2001). Application of multidimensional scaling to subjective evaluation of coded speech. *The Journal of the Acoustical Society of America*, *110*(4), 2167–2182. https://doi.org/10.1121/1.1397322.

Hansakunbuntheung, C., Rugchatjaroen, A., & Wutiwiwatchai, C. (2005). Space reduction of speech corpus based on quality perception for unit selection speech synthesis. In *Proceedings of the 6th international symposium on natural language processing* (pp. 127–132). https://www.researchgate.net/profile/Chatchawarn_Hansakunbuntheung/publication/228957899_Space_reduction_of_speech_corpus_based_on_quality_perception_for_unit_selection_speech_synthesis/links/0912f510bb45091b12000000.pdf.

Harris, J. (1983). *Syllable structure and Stress in Spanish*. Cambridge: The MIT Press.

Hinterleitner, F., Norrenbrock, C., & Möller, S. (2013). Is intelligibility still the main problem? A review of perceptual quality dimensions of synthetic speech. In *Proceedings of the eighth ISCA workshop on speech synthesis* (pp. 147–151). http://ssw8.talp.cat/papers/ssw8_PS2-1_Hinterleitner.pdf.

Hinterleitner, F., Norrenbrock, C., Möller, S., & Heute, U. (2014). *Text-to-speech synthesis. Quality of experience* (pp. 179–193). Berlin: Springer.

Hinterleitner, F., Zabel, S., Möller, S., Leutelt, L., & Norrenbrock, C. (2011). Predicting the quality of synthesized speech using reference-based prediction measures. In *Proceedings of the 22th Konferenz Elektronische Sprachsignalverarbeitung* (pp. 99–106). http://www.qu.tu-berlin.de/fileadmin/fg41/publications/hinterleitner_2011_predicting-the-quality-of-synthesized-speech-using-reference.-.based-prediction-measures.pdf.

Hirst, D., Rilliard, A., & Aubergé, V. (1998). Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis. In *Proceedings of the third ESCA/COCOSDA workshop (ETRW) on speech synthesis* (pp. 293–306). https://www.isca-speech.org/archive_open/archive_papers/ssw3/ssw3_001.pdf.

Hoeckel, C. (1989). The reliability of manual labelling of continuous speech. In *Proceedings of the ESCA workshop on speech input/output assessment an speech databases* (Vol. 2, pp. 2179–2182). http://www.isca-speech.org/archive_open/archive_papers/sioa_89/sia_2179.pdf.

Hon, H., Acero, A., Huang, X., Liu, J., & Plumpe, M. (1998). Automatic generation of synthesis units for trainable text to speech systems. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP'98)* (Vol. 1, pp. 293–306). https://doi.org/10.1109/ICASSP.1998.674425

Karabetsos, S., Tsiakoulis, P., Chalamandaris, A., & Raptis, S. (2009). Embedded unit selection text-to-speech synthesis for mobile devices. *IEEE Transactions on Consumer Electronics*, *55*(2), 613–621. https://doi.org/10.1109/TCE.2009.5174430.

Kawai, H., & Toda, T. (2004). An evaluation of automatic phone segmentation for concatenative speech synthesis. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing* (Vol. 1, pp. I–677–80). https://doi.org/10.1109/ICASSP.2004.1326076.

Kawai, H., & Tsuzaki, M. (2002). Study on time-dependent voice quality variation in a large-scale single speaker speech corpus used for speech synthesis. In *Proceedings of the IEEE workshop on speech synthesis* (pp. 15–18). https://doi.org/10.1109/WSS.2002.1224362.

Kelly, A. C., Berthelsen, H., Campbell, N., Chasaide, A. N., & Gobl, C. (2009). Corpus design techniques for irish speech synthesis. In *Proceedings of the China Ireland ICT conference* (pp. 264–265). http://www.eeng.dcu.ie/ciict/2009/proceedings.pdf.

King, S. (2014). Measuring a decade of progress in text-to-speech. *Loquens, 1*(1). https://doi.org/10.3989/loquens.2014.006.

Kishore, S., & Black, A. (2003). Unit size in unit selection speech synthesis. In *Proceedings of the Eurospeech 2003* (pp. 1317–1320). https://www.isca-speech.org/archive/archive_papers/eurospeech_2003/e03_1317.pdf.

Krul, A., Damnati, G., Yvon, F., Boidin, C., & Moudenc, T. (2007). Approaches for adaptive database reduction for text-to-speech synthesis. In *Proceedings of the eighth annual conference of the international speech communication association* (Vol. 3, pp. 2881–2884). https://www.isca-speech.org/archive/archive_papers/interspeech_2007/i07_2881.pdf.

Kurtic, E. (2004). Polyglot voice design for unit selection speech synthesis. Master's thesis, School of Philosophy, Psychology and Language Sciences, University of Edinburgh. https://www.era.lib.ed.ac.uk/bitstream/handle/1842/2070/Emina%20Kurtic.pdf?sequence=1&isAllowed=y

Lambert. T., Braunschweiler, N., & Buchholz, S. (2007). How (not) to select your voice corpus: Random selection vs. phonologically balanced. In *Proceedings of the 6th ISCA workshop on speech synthesis* (pp. 22–24). https://isca-speech.org/archive_open/archive_papers/ssw6/ssw6_264.pdf.

Lewis, E., & Tatham, M. (1999). Word and syllable concatenation in text-to-speech synthesis. In *Proceedings of the sixth European conference on speech communications and technology* (Vol. 2, pp. 615–618). https://www.isca-speech.org/archive/archive_papers/eurospeech_1999/e99_0615.pdf.

Llisterri, J. (1999). Transcripción, etiquetado y codificación de corpus orales. *Revista Española de Lingüística Aplicada*, Monográfico: Panorama de la Investigación en Lingüística Informática, (pp, 53–82). http://liceu.uab.es/~joaquim/publicacions/RESLA_99.pdf.

Lu, H., Zhang, W., Shao, X., Lei, Q. Z. W., Zhou, H.. & Breen, A. (2015). Pruning redundant synthesis units based on static and delta unit appearance frequency. In *Proceedings of the sixteenth annual conference of the international speech communication association* (pp. 269–273). https://www.isca-speech.org/archive/interspeech_2015/papers/i15_0269.pdf.

Marino, J. B., Nogueiras, A., Pachès-Leal, P., & Bonafonte, A. (2000). The demiphone: An efficient contextual subword unit for continuous speech recognition. *Speech Communication*, *32*(3), 187–197. https://doi.org/10.1016/S0167-6393(00)00010-8.

Matoušek, J., & Psutka, J. (2001). Design of speech corpus for text-to-speech synthesis. In *Proceedings of the 7th conference on speech communication and technology* (pp. 2047–2050). https://www.isca-speech.org/archive/archive_papers/eurospeech_2001/e01_2047.pdf.

Matoušek, J., Tihelka, D., & Romportl, J. (2008). Building of a speech corpus optimised for unit selection TTS synthesis. In *Proceedings of 6th international conference on language resources and evaluation* (pp. 1296–1299). http://www.lrec-conf.org/proceedings/lrec2008/pdf/329_paper.pdf.

Mayo, C., Clark, R. A., & King, S. (2005). Multidimensional scaling of listener responses to synthetic speech. In *Proceedings of the 9th European conference on speech communication and technology* (pp. 1725–1728). https://www.isca-speech.org/archive/archive_papers/interspeech_2005/i05_1725.pdf.

McPherson, I. (1975). *Spanish phonology: Descriptive and historical*. Manchester: Manchester Univiversity Press.

Mendelson, J., & Aylett, M. (2017). Beyond the listening test: An interactive approach to TTS evaluation. In *Proceedings of the 18th annual conference of the international speech communication association* (pp. 20–24). https://doi.org/10.21437/Interspeech.2017-1438.

Möbius, B. (2000). Corpus-based speech synthesis: Methods and challenges. *AIMS, Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung, 6*(4), 87–116. http://www.ims.uni-stuttgart.de/~moebius/papers/unitsel.pdf.

Möbius, B. (2003). Rare events and closed domains: Two delicate concepts in speech synthesis. *International Journal of Speech Technology*, *6*(1), 57–71. https://doi.org/10.1023/A:1021052023237.

Möller, S., Hinterleitner, F., Falk, T. H., & Polzehl, T. (2010). Comparison of approaches for instrumentally predicting the quality of text-to-speech systems. In *Proceedings of the eleventh annual conference of the international speech communication association* (pp. 1325–1328). https://www.isca-speech.org/archive/archive_papers/interspeech_2010/i10_1325.pdf.

Ni, J., Hirai, T., Kawai, H., Toda, T., Tokuda, K., Tsuzaki, M., Sakai, S., Maia, R., & Nakamura, S. (2007). ATRECSS: ATR english speech corpus for speech synthesis. In *Proceedings of the 6th ISCA workshop on speech synthesis, paper 002*. https://www.isca-speech.org/archive_open/archive_papers/blizzard_2007/blz3_002.pdf.

Niebuhr, O., & Michaud, A. (2015). Speech data acquisition: The underestimated challenge. In *KALIPHO-Kieler Arbeiten zur Linguistik und Phonetik, 3,* 1–42. https://halshs.archives-ouvertes.fr/halshs-01026295v4/document.

Norrenbrock, C. R., Hinterleitner, F., Heute, U., & Möller, S. (2015). Quality prediction of synthesized speech based on perceptual quality dimensions. *Speech Communication*, *66*, 17–35. https://doi.org/10.1016/j.specom.2014.06.003.

Oliveira, L. C., Paulo, S., Figueira, L., Mendes, C., Nunes, A., & Godinho, J. (2008). Methodologies for designing and recording speech databases for corpus based synthesis. In *Proceedings of the 6th international conference on language resources and evaluation* (pp. 2921–2925). http://www.lrec-conf.org/proceedings/lrec2008/pdf/741_paper.pdf.

P.85 ITR. (1990). Studies toward the unification of picture assessment methodology. *Technical report, ITU*. https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-BT.1082-1-1990-PDF-E.pdf.

P800 ITR. (1996). Methods for subjective determination of transmission quality. *Technical report, ITU*. https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-P.800-199608-I!!PDF-E&type=items.

P85 ITR. (1994). Method for subjective performance assessment of the quality of speech voice output devices. *Technical report, ITU*. https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-P.85-199406-I!!PDF-E&type=items.

Peterson, G. E., Wang, W. S. Y., & Sivertsen, E. (1958). Segmentation techniques in speech synthesis. *The Journal of the Acoustical Society of America*, *30*(8), 739–742. https://doi.org/10.1121/1.1909746.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89–95. https://doi.org/10.1016/j.specom.2004.09.001.

Prudon, R., & d'Alessandro, C. (2001). A selection/concatenation text to speech synthesis system: Databases development, system design, comparative evaluation. In *Proceedings of the 4th speech*

synthesis workshop (SSW4-2001), paper 138. https://www.isca-speech.org/archive_open/archive_papers/ssw4/ssw4_138.pdf.

Rodríguez, H. (2000). Construcción de una base de datos para el desarrollo de sistemas de conversión de texto a habla. University of La Plata, Buenos Aires, licenciature thesis.

Rosenberg, A., & Ramabhadran, B. (2017). Bias and statistical significance in evaluating speech synthesis with mean opinion scores. In *Proceedings of the 18th annual conference of the international speech communication association* (pp. 3976–3980). https://doi.org/10.21437/Interspeech.2017-479.

Royal Spanish Academy. (1992). *Dictionary of the Spanish language*. Madrid: Espasa Calpe.

Rutten, P., Aylett, M. P., Fackrell, J., & Taylor, P. (2002). A statistically motivated database pruning technique for unit selection synthesis. In *Proceedings of the seventh international conference on spoken language processing* (pp. 125–128). https://www.isca-speech.org/archive/archive_papers/icslp_2002/i02_0125.pdf.

Sainz, I., Navas, E., Hernáez, I., Bonafonte, A., & Campillo, F. (2010). TTS evaluation campaign with a common spanish database. In *Proceedings of the seventh international conference on language resources and evaluation* (pp. 2155–2160). http://www.lrec-conf.org/proceedings/lrec2010/pdf/456_Paper.pdf.

Schiel, F., Baumann, A., Draxler, C., Ellbogen, T., Hoole, P., & Steffen, A. (2012). The validation of speech corpora. Munchen: Bavarian Archive for Speech Signals. https://epub.ub.uni-muenchen.de/13698/1/schiel_13698.pdf.

Sityaev, D., Knill, K., & Burrows, T. (2006). Comparison of the ITU-T P.85 standard to other methods for the evaluation of text-to-speech systems. In *Proceedings of the ninth international conference on spoken language processing* (pp. 2743–2746). https://www.isca-speech.org/archive/archive_papers/interspeech_2006/i06_1233.pdf.

Streijl, R. C., Winkler, S., & Hands, D. S. (2016). Mean opinion score (mos) revisited: Methods and applications, limitations and alternatives. *Multimedia Systems*, *22*(2), 213–227. https://doi.org/10.1007/s00530-014-0446-1.

Syrdal, A., Wightman, C., Conkie, A., Stylianou, Y., Beutnagel, M., Schroeter, J., Strom, V., Lee, K., & Makashay, M. (2000). Corpus-based techniques in the AT&t nextgen synthesis system. In *Proceedings of the 6th international conference on spoken language processing* (Vol. 3, pp. 410–415). https://www.isca-speech.org/archive/archive_papers/icslp_2000/i00_3410.pdf.

Syrdal, A. K., Conkie, A., & Stylianou, Y. (1998). Exploration of acoustic correlates in speaker selection for concatenative synthesis. In *Proceedings of the international conference on spoken language processing* (Vol. 6, pp. 2743–2746). https://www.isca-speech.org/archive/archive_papers/icslp_1998/i98_0882.pdf.

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge: Cambridge University Press.

Torres, H. M. (2012). Creación de un corpus de texto para la construcción de un sistema TTS. *Informe técnico*, ISSN 0325-2043, Laboratorio de Investigaciones Sensoriales, UBA-CONICET, Buenos Aires, Argentina. http://www.lis.secyt.gov.ar/informes/2012.pdf

Torres, H. M. (2013). Medición de la velocidad de conversión del sistema TTS aromo. *Informe técnico*, ISSN 0325-2043, Laboratorio de Investigaciones Sensoriales, UBA-CONICET, Buenos Aires, Argentina. http://www.lis.secyt.gov.ar/informes/2013.pdf

Torres, H. M., & Gurlekian, J. (2004). Automatic determination of phrase breaks for argentine spanish. In *Proceedings of the speech prosody 2004* (pp. 553–556). http://www.isca-speech.org/archive_open/sp2004/sp04_553.pdf.

Torres, H. M., & Gurlekian, J. A. (2008). Acoustic speech unit segmentation for concatenative synthesis. *Computer Speech and Language*, *22*, 196–206. https://doi.org/10.1016/j.csl.2007.07.002.

Torres, H. M., & Gurlekian, J. A. (2009). Parameter estimation and prediction from text for a superpositional intonation model. In *Proceedings of the 20 Konferenz Elektronische Sprachsignalverarbeitung* (pp. 238–247). https://www.researchgate.net/publication/265963364_Parameter_estimation_and_prediction_from_text_for_a_superpositional_intonation_model

Torres, H. M., & Gurlekian, J. A. (2016). Novel estimation method for the superpositional intonation model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(1), 151–160. https://doi.org/10.1109/TASLP.2015.2500728.

Torres, H. M., Gurlekian, J. A., & Mercado, C. (2012). Aromo: Argentine spanish TTS system. In *Proceedings of VII Jornadas en Tecnología del Habla and III Iberian SLTech workshop* (pp. 416–421). https://www.researchgate.net/profile/Christian_Cossio-Mercado/publication/265952108_Aromo_Argentine_Spanish_TTS_System/links/570c37ea08aee0660351b0b9.pdf

Umbert, M., Moreno, A., Agüero, P., & Bonafonte, A. (2006). Spanish synthesis corpora. In *Proceedings of the international conference of language resources and evaluation* (pp. 2102–2105). http://www.lrec-conf.org/proceedings/lrec2006/pdf/590_pdf.pdf.

Vainio, M., Jarvikivi, J., Werner, S., Volk, N., & Valikangas, J. (2002). Effect of prosodic naturalness on segmental acceptability in synthetic speech. In *Proceedings of 2002 IEEE workshop on speech synthesis* (pp. 143–146). https://doi.org/10.1109/WSS.2002.1224394.

Valentini-Botinhao, C., Yamagishi, J., & King, S. (2011). Evaluation of objective measures for intelligibility prediction of HMM-based synthetic speech in noise. In *2011 IEEE international conference on acoustics, speech and signal processing* (pp. 5112–5115). https://doi.org/10.1109/ICASSP.2011.5947507.

van den Heuvel, H., Iskra, D., Sanders, E., & de Vriend, F. (2008). Validation of spoken language resources: An overview of basic aspects. *Language Resources and Evaluation*, *42*(1), 41–73. https://doi.org/10.1007/s10579-007-9049-1.

van Santen, J. P. H. (1997). Prosodic modelling in text-to-speech synthesis. In *Proceedings of the 5th European conference on speech communication and technology* (Vol. 5, pp. 2511–2514). https://www.isca-speech.org/archive/archive_papers/eurospeech_1997/e97_KN19.pdf.

Viswanathan, M., & Viswanathan, M. (2005). Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (mos) scale. *Computer Speech & Language*, *19*(1), 55–83. https://doi.org/10.1016/j.csl.2003.12.001.

Watson, A., Mullin, J., Smallwood, L., & Wilson, G. (2001). New techniques for assessing audio and video quality in real-time interactive communication. In *Tutorial at IHM-HCI*, Lille, France. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.494.6094&rep=rep1&;type=pdf.

Zhang, W., Liu, Y., Deng, Y., & Pang, M. (2010). Automatic construction for a TTS corpus with limited text. In *Proccedings of the 2010 international conference on measuring technology and mechatronics automation* (Vol. 1, pp. 707–710). https://doi.org/10.1109/ICMTMA.2010.796.