ON MIRANDESE LANGUAGE RESOURCES FOR TEXT-TO-SPEECH

José Pedro Ferreira[2], *Cristiano Chesi*[1,3,4], *Hyongsil Cho* [1, 3], *Daan Baldewijns*[1], *Daniela Braga*[1, 3], *Miguel Dias* [1, 3]

[1]Microsoft Language Development Center, Portugal.
[2]Instituto de Linguística Teórica e Computacional. [3]ISCTE-IUL University Institute of Lisbon,
[4]IUSS, Istituto Universitario di Studi Superiori, Pavia

ABSTRACT

This paper aims at describing the major components of the first Text-to-Speech (TTS) system ever built for Mirandese, [1] a minority language spoken in the Northeast of Portugal. Both language resources development (corpus, text-normalization rules, annotated lexicon, phone sets and recordings) and the TTS (Statistical Parameter Synthesis) system are documented here.

Index Terms**—** Mirandese, text-to-speech, language resources

1. INTRODUCTION

Mirandese is a minority language spoken by around 15.000 people in the Northeast of Portugal for which little or no computational resources were available before the ones described in this paper. To pursue the development of a Mirandese Text-to-Speech system, Microsoft teamed up with a public research center, ILTEC, and with the community-led Association for Mirandese Language (ALM). With the goal of building the first TTS system, language resources for Mirandese have been developed, consisting of:

i. a corpus of over 1 million tokens;
ii. a set of about 4.000 contextual text normalization rules;
iii. tokenizer, inflector, stress maker, syllabifier and grapheme-to-phoneme converter;
iv. a lexicon fully annotated for Part-of-Speech and other morpho-syntactic features (124.360 word forms);
v. a phone set including 46 distinct phones;
vi. a speech data base composed of about 7 hours of recordings from a carefully selected voice talent, using about 5.000 prompts retrieved from the corpus.

After a brief introduction to Mirandese language (§2), these resources are described (§3), and the TTS system presented (§4). An overall evaluation of the quality of the modules developed so far concludes this paper (§5).

2. THE MIRANDESE LANGUAGE

Mirandese belongs to the Astur-Leonese group of West Iberian languages, being closely related to Asturian, spoken to this day in areas of the Asturias and Leon autonomous communities in Spain, with which Mirandese no longer retains a linguistic continuum. Unwritten for most of its history, Mirandese was first scientifically identified and studied in the late 19th century [2]. Throughout the 20th century, strong demographic changes, namely the exodus of large numbers through emigration in the 1940s and in the 1970s, an influx of non-Mirandese speaking workers from various other parts of the country in the 1950s and 1960s, along with the rise of Portuguese-spoken-only media, led to inter-generational transmission to be gradually abandoned, leaving the language seriously threatened. Today, it is estimated that Mirandese is spoken by no more than 5.000 people as a first language, and by at most 15.000 in total, counting heritage and second language speakers, including those living outside of Terra de Miranda [3]. In the 1990s, strong efforts began to be made to make the survival of Mirandese possible: a group of linguists and native speakers managed to reach an agreement for a spelling convention common to different varieties, and the language was introduced into the formal education curricula locally, although with limited scope and only as an option. The Portuguese State finally granted the language co-official regional status in 1999 [4]. These initiatives had a strong impact on its speakers: what used to be perceived as a reason for shame by many in the diglossic community increasingly started showing up in book shelves and in the local and national media, and on the Web. Albeit seriously threatened as a mother tongue, it is currently learned and used by a large part of the population of Miranda in increasingly more formal contexts, currently enjoying a period of non-artificial revival.

2.1. Some properties of Mirandese phonetics

Mirandese is closely related to other central and western Iberian romance languages. Although the matter is not definitively settled, it features five distinctive vowels in the stressed system (/i, e, a, o, u/), like Spanish, but nine corresponding phones if unstressed positions are taken into account ([i, e, ɛ, ɯ, ɐ, a, u, o, ɔ]), like Portuguese.

It also shares with this language the existence of nasal vowels (including nasal diphthongs and the rare nasal schwa), although with different height values; unlike it, it has a high number of rising diphthongs, a feature again more similar to Spanish and to other, closely related Asturian-Leonese dialects, with which it also shares word-initial palatalization of laterals. Among its most distinctive features is the phonological differentiation of 7 sibilants: apico-alveolar /s̺/ and /z̺/, laminal / predorso-dental /s/ and /z/, along with /ʃ/, /ʒ/, and the africate [tʃ].

On a phonetic level, some Mirandese varieties feature systematic diphthongation of stressed /i,u/, which become glides attached to a near-back nucleous; systematic palatalization or elision of velar stops before a stressed close front oral or nasal vowel, and contextual intervocalic velar nasal [ŋ] after a stressed nasal /u/ preceding a vowel.

3. LANGUAGE RESOURCES

Despite some pioneering efforts towards developing speech technologies for Mirandese [6] [7], there were little or no available resources at the onset of this effort [1]. The most detailed linguistic descriptions are those made by [4], more than 100 years ago, in part due to the lack of available data [3], leaving researchers in need of conducting original fieldwork to get in touch with actual large-scale data, and school pupils with little up-to-date base tools for the formal study of the language. Additionally, the fact that Mirandese is present in more and more support formats and usage contexts seems to be a decisive factor in the way its speakers perceive the language, granting it a higher sociolinguistic profile [3].

To achieve the end goal of creating a TTS system, a number of language resources that are usually available for more widely spoken European languages had to be developed from scratch. For the voice font generation, an existing and proven process could be followed, using previously existing tools designed for larger and better-resourced languages [5] [8]. Work for this project encompassed the creation of a large text corpus, the definition of a complete phone-set, the development of tokenizer, inflector, text normalization (TN) and grapheme-to-phoneme (GTP) modules, and the creation of a large phonetic lexicon with part-of-speech (POS) classification.

3.1. Corpus

The corpus was compiled from raw textual data collected by ALM, most of it generously provided by publishers, newspapers and the authors themselves. Those data were then complemented with data crawled from the web using the work developed by [9], which allowed us to harvest authentic data from sources such as blog posts and comments and personal websites, increasing the total size of the corpus to over one million tokens.

The data in the corpus was used to retrieve the text prompts used in the recordings, extract lexical entries, and to test the Text Normalization (TN) module.

3.2. TN module

A TN module for Mirandese was also put in place for the first time for this language, its rule-set being developed with the aid of previously in-house developed software [15] and taking advantage of the availability of a counterpart file for European Portuguese. The proximity between the two languages made it possible to reuse the pre-existing resource changing only minimally several hundreds of the thousands of rules and terminals that compose the module, thus greatly speeding up the TN module development process.

The final TN module is composed of about 5.000 (contextual) rules dealing with the expansion of cardinal numbers (“12” ↔ “twelve”) or date-time spell out (“12:00” ↔ “noon”), for instance. The following TN categories were developed for Mirandese: cardinals, ordinals, percentage expressions, simple mathematical expressions, date and time expressions, currency, phone numbers, roman numerals, fractions, measurement expressions, titles, addresses, URLs and email, and file paths. Where needed, context-sensitive rules were made, for instance to ensure the correct gender agreement in noun phrases containing a cardinal number.

3.3. Lexicon

The lexicon was built using the currently 25K lemma list of the ongoing work on a Mirandese-Portuguese dictionary [10], complemented with the most frequent lemmas in the compiled text corpus, which was previously tokenized, lemmatized and POS-tagged using customizations of [11] and [12]. The resulting lemma list was then inflected using a version of [13], syllabified, stress-marked and converted to IPA phonetic transcription using an in-house adaptation of an unpublished two-step Perl-based GTP tool originally developed for European Portuguese [14]. This simple regular expression string replacement set of scripts starts by marking up stress and syllable divisions over the orthographic forms and, in a subsequent phase, applies an ordered set of grapheme-phone transformation rules based on syllable position and stress, taking advantage of the relatively shallow phonemic orthography of Mirandese.

In the end, we succeeded in compiling a fully annotated large lexicon, consisting of 124.360 word forms, for which standard orthography, pronunciation (syllabified, stress marked IPA transcription), POS (e.g. VER for verb, ADJ for adjective) as well as other morphological information (like mood, tense, person and number features) is provided, as exemplified in (1):

(1) *Word | Pronunciation | POS | morphological features*
abacelhe | ax - b ax - s eh 1 - lh aex | VER |
subjunctive, present, 3person, sing
melhor | m aex - lh oh r 1 | ADJ | qualifying, masc, sing

Using algorithms like the one discussed in [14], we managed to easily generate syllabification rules that allow us to segment words out of lexicon and pair them with their correct pronunciation and the most likely stress.

3.4. Phone set

Another crucial resource that needed to be developed, was a complete phone set for Mirandese (cf. §2.1), consisting of 43 distinct phones, which that takes us forward from prior work [6]. This resource specifies the full list of available phones paired to distinctive features and parameters that are used to train the voice model:

(2)

Phone	*features*
a	voiced central low sonorant vowel
aex	voiced central high sonorant vowel
an	voiced central low sonorant nasal vowel
ax	voiced central mid-high sonorant vowel
b	voiced bilabial plosive
ch	voiceless postalveolar affricate
d	voiced dental plosive
eh	voiced front mid-high sonorant vowel
ehn	voiced front mid-low sonorant nasal vowel
exn	voiced central high sonorant nasal vowel
f	voiceless labiodental fricative
g	voiced velar plosive
i	voiced front high sonorant vowel
in	voiced front high sonorant nasal vowel
j	voiced palatal semivowel
je	rising front high sonorant diphthong
jen	rising front high sonorant nasal diphthong
k	voiceless velar plosive
l	voiced alveolar lateral sonorant
lg	voiced lateral velarized alveolar sonorant
lh	voiced palatal lateral sonorant
m	voiced bilabial nasal sonorant
n	voiced alveolar sonorant nasal
ng	voiced velar nasal sonorant
nh	voiced palatal nasal sonorant
oh	voiced back mid-high rounded sonorant vowel
ohn	voiced back mid-low rounded sonorant nasal vowel
p	voiceless bilabial plosive
r	voiced alveolar tap
rr	voiced alveolar trill
s	voiceless alveolar fricative
sh	voiceless postalveolar fricative
ss	voiceless apicoalveolar fricative
t	voiceless dental plosive
u	voiced high rounded sonorant vowel
un	voiced high rounded sonorant nasal vowel
w	voiced labiovelar semivowel
wo	rising back high sonorant diphthong
won	rising back high sonorant nasal diphthong
x	voiceless uvular fricative
z	voiced alveolar fricative
zh	voiced postalveolar fricative
zz	voiced apicoalveolar fricative

3.5. Voice recordings

On par with the other development tasks, a voice talent was selected for high-quality recording sessions of 5.132 prompts retrieved from the corpus. The prompts consisted of full sentences selected based on character length and phonological relevance (richness of phonological contexts), determined by an existing algorithm of the text TTS system software suite.

The voice talent was selected from a pool of candidates by a jury of 20 native speakers of varying ages, provenances and sociolinguistic profiles. Public advertisement in the local media and through speaking community networking helped greatly in getting a reasonable number of candidates with the correct profile: native speakers, having at least undergone undergraduate studies and no older than 40. The jury listened to recordings of each candidate reading an expressive text, and filled in a short questionnaire. The two highest ranked candidates in this first phase underwent one hour of pure speech studio recording beneath loose scrutiny and were again ranked by the jury, who this time had to fill in a more thorough questionnaire developed for subjective pleasantness assessment, using a methodology published before [17].

Finally, the selected voice talent was recorded over two weeks in a high-quality studio under the close supervision of a Language Expert, who monitored the clarity, accent, and completeness of the recording process, simultaneously checking the adequacy of each prompt and making textual corrections where needed. The recording process yielded over 7 hours of speech data.

Those data were semi-automatically trimmed and chopped into individual files using a standard acoustic marker inserted between prompts during the recording process, making it easier to map each individual recording file with a prompt. All the individual recordings were listened to by a Language Expert, and removed from the pool of available data when quality or conformity with the prompt was not met.

4. TEXT-TO-SPEECH SYSTEM

The resources developed have been used for the Voice Font Building procedure at the core of the TTS system discussed in [1]. In this section we quickly discuss the TTS back-end voice font training procedure.

The Statistical Parameter Synthesis (SPS) used requires the extraction of (up to) 24 parameters characterizing the vocal tract emitting the relevant phone to be modeled. The wave segmentation (at sentence, word, syllable and phone level) is guided by the fully normalized prompts used for eliciting the recording. This procedure is automatically carried out using rule-based sentence breakers, contextual text normalization (TN) rules (as discussed in §3.2), and POS taggers (e.g. [18]). Once single words are normalized and categorized, the correct pronunciation is retrieved from the lexicon (§3.3) and assigned to the current word.

In the end, each prompt is enriched with several types of information, as shown in (3) ("w"= token, "v"=written form, "p"= pronunciation):

(3) Por baixo de las saias de las rapazas
in under of the skirts of the ladies

```
<w v="Por" p="p . oh 1 . r" type="normal"
        length="3" />
<w v="baixo" p="b . a 1 . j - ch . u" type="normal"
        offset="4" length="5" />
<w v="de" p="d . aex" type="normal" offset="10"
        length="2" />
<w v="las" p="l . ax . ss" type="normal" offset="13"
        length="3" />
<w v="saias" p="ss . a 1 . j - ax . ss" type="normal"
        offset="17" length="5" />
<w v="de" p="d . aex" type="normal" offset="23"
        length="2" />
<w v="las" p="l . ax . ss" type="normal" offset="26"
        length="3" />
<w v="rapazas" p="rr . ax - p . a 1 - z . ax . ss"
        type="normal" br="4" offset="30"
        length="7" />
<w v="." type="punc" br="4" />
```

The fully annotated prompts made the TTS voice font training procedure possible: the approach used to train the font model is called Statistical Parameter Synthesis (SPS) and it is based on standard Hidden-Markov-Model (HMM) approaches to TTS (HTS, [19], [20]).

Notice that the SPS procedure not only allows using distinctive phonetic features (linear spectrum pair, LSP model, [21]) as parameters, but also prosodic cues, like pitch (F0). This allows us to both keep the advantages of having an HTS Voice Font (high flexibility, small font size) and to limit its disadvantages (muffle voice quality, flat prosody). The training process uses a gradient descent algorithm (Minimum Generation Error, MGE [22]).

In the end, the (trained) decision tree is used in generation to select and concatenate the state models by maximizing the likelihood of the parameter sequence. The result is a fully intelligible TTS voice font.

5. CONCLUSIVE REMARKS

In this work, we discussed the linguistic resources developed for the first TTS system for Mirandese reaching intelligibility. The language resources developed to build the system are being made available to the speaking and scientific community through the speech-community-led Casa de la Lhéngua (free access web interface), as well as through the Microsoft Language Development Center web site. It rests to be seen if the secondary objective of granting Mirandese a stronger sociolinguistic profile within its speaking community will be aided by the availability of these tools.

Future work should include the conversion of the NLP resources we developed to internationally standardized formats and the general improvement of the TTS system. We are now working on a more natural TTS voice font, where the prosody model used will be improved using an enriched TOBI-compliant prompt annotation. [23] With an evaluation purpose, we will conduct in the near future a community-centered evaluation of the TTS system, namely at the segmental level (including intelligibility tests), as well as at supra-segmental, prosodic level, by means of Mean Opinion Scores (MOS). [24].

6. REFERENCES

[1] J. P. Ferreira, C. Chesi, D. Baldewijns, D. Braga, M.S. Dias 2012. *The First Mirandese Text-to-Speech System*. *Proceedings of ELE2013 Conference*.

[2] Branco, A., Mendes, A., Pereira, S., Henriques, P., Pellegrini, T., Meinedo, H., Trancoso, I., Quaresma, P. S. de Lima, V. L. 2012. *The Portuguese Language in the Digital Age*. Berlin: Springer.

[3] Barros Ferreira, M. 2002. "O mirandês, língua minoritária". In Mira Mateus, M. H. (org.), *Uma política de língua para o português*. Lisbon: Colibri: 137-145.

[4] Leite Vasconcelos, J. 1899-1900[1990]. *Lições de Filologia Mirandesa*. Miranda do Douro: Câmara Municipal de Miranda do Douro.

[5] Braga, D., Campillo, F., Dias, M.S., García-Mateo, C., Méndez, F., Mourín, A., Silva, P. 2010. "Building high quality databases for minority languages such as Galician". In Calzolari, N. et al. (eds.), *Proc. of LREC'10*. La Valletta: ELRA:113-116.

[6] Trancoso, I., Ribeiro, V., Barros, M., Caseiro, A., D., Paulo, S., "From Portuguese to Mirandese: fast porting of a letter-to-sound module using FSTs". In Mamede, N. J. et al. (eds.), *Proc. of PROPOR'2003*. LNCS. Berlin / Heidelberg: Springer: 49-56.

[7] Caseiro, A., D., Trancoso, I., Guerreiro, M., C., Ribeiro, V., Barros, M., "A Comparative Description of GtoP modules for Portuguese and Mirandese using Finite State Transducers". In In Solé et al. (eds.), *Proc. of ICPhS'2003*: 2605-2608.

[8] Braga, D., Silva, P., Ribeiro, M., Henriques, M. and Dias, M. 2008. "HMM-based Brazilian Portuguese TTS". In Braga et al. (eds), *Propor 2008 Special Session: Applications of Portuguese Speech and Language Technologies, September 10, 2008, Curia, Portugal*: 47-50. Available at http://download.microsoft.com/download/A/0/B/A0B1A66A-5EBF-4CF3-9453-4B13BB027F1F/Braga_Propor08.pdf, retr. 13 Jan 2014.

[9] Scannell, K. 2007. "The Crúbadán Project: Corpus building for under-resourced languages". In Fairon et al. (eds.), *Building and Exploring Web Corpora, Proceedings of the 3rd Web as Corpus Workshop* (*Cahiers du Cental 4*). Louvain-la-Neuve: Université Catholique de Louvain, 5-15.

[10] Ferreira, A., Ferreira, J. P. 2001-. *Dicionário Mirandês-Português*. Lisbon: authors. Available at http://www.mirandadodouro.com/dicionario/, retr. 13 Jan. 2014.

[11] García, M., Gamallo, P. 2010. "Análise Morfossintáctica para o Português Europeu e Galego: Problemas, Soluções e Avaliação". *Linguamática*, 2(2): 59-67.

[12] Janssen, M. 2012. "NeoTag: a POS Tagger for Grammatical Neologism Detection". In Calzolari, N. et al. (eds.), *Proc. of LREC'12*. Istanbul: ELRA: 2118-2124.

[13] Janssen, M. 2011. "Computer-Aided Inflection for Lexicography Controlled Lexica". In Kosem, I., Kosem, K. (Eds.). *Proceedings of eLex 2011*. Ljublijana: Trojína: 96-105.

[14] Janssen, M., Santos, F. 2012. "Building a database of phonetic transcriptions from a speech corpus". *VII GSCP International Conference: Speech And Corpora, 29 Februry – 3 March 2012, Belo Horizonte, Brazil.*

[15] Cho, H, Braga, D., Chesi, C., Baldewijns, D., Ribeiro, M., Saarinen, K., Beck, J., Rustullet, S., Henriksson, P, Dias, M., Rahmel, H. 2010. "A Multi-lingual TN/ITN Framework for Speech Technology". In García Mateo, C., Campillo Díaz, F., Méndez Pazó, F. (eds.), *Proceedings of FALA 2010*. Vigo, Universidad de Vigo: 213-216.

[16] Müller, K., Möbius, B., Presher D. 2000. "Inducing probabilistic syllable classes using multivariate clustering". In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*: 225-232.

[17] Braga, D., Coelho, L., Resende, F. G., Dias, M. 2008. "Subjective and Objective Evaluation of Brazilian Portuguese TTS Voice Font Quality". In Kacic, Z. and Markus, A. (eds.), *Advances in Speech Technology – Proceedings of the 14th International Workshop*. Maribor: Faculty of Electrical Engineering and Computer Science: 129-138.

[18] Ratnaparkhi, A. 1996. "A maximum entropy model for part-of-speech tagging". In *Proceedings of the Empirical Methods in Natural Language Processing 1*. New Brunswick, New Jersey: ACL: 133-142.

[19] Zen, H., T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black and K. Tokuda. 2007. "The HMM-based speech synthesis system version 2.0". In *Speech Synthesis Workshop, Bonn, Germany*: 294-299.

[20] Zen, H., Tokuda, K., & Black, A. W. 2009. "Statistical parametric speech synthesis". *Speech Communication*, 51(11): 1039-1064.

[21] Zheng F., Z. Song, W. Yu, F. Zheng, W. Wu. 2000. "The distance measure for line spectrum pairs applied to speech recognition". *Journal of Computer Processing of Oriental Languages*, 11: 221-225.

[22] Yi-Jian Wu, and Ren-Hua Wang. 2006. "Minimum generation error training for HMM-based speech synthesis". In *Proc. of ICASSP*: 89-92.

[23] Beckman, M. E., & Hirschberg, J. 1994. *The ToBI annotation conventions*. Manuscript, Ohio State University.

[24] 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.