Building Synthetic Voices in the METANET Framework

Emília Garcia-Casademont, Antonio Bonafonte, Asunción Moreno

Universitat Politècnica de Catalunya

Abstract

METANET$_4$U is a European project aiming at supporting language technology for European languages and multilingualism. It is a project in the META-NET Network of Excellence, a cluster of projects aiming at fostering the mission of META, which is the Multilingual Europe Technology Alliance, dedicated to building the technological foundations of a multilingual European information society. This paper describe the resources produced at our lab to provide Synthethic voices. Using existing 10h corpus for a male and a female Spanish speakers, voices have been developed to be used in Festival, both with unit-selection and with statistical-based technologies. Furthermore, using data produced for supporting research on intra and inter-lingual voice conversion, four bilingual voices (English/Spanish) have been developed. The paper describes these resources which are available through META. Furthermore, an evaluation is presented to compare different synthesis techniques, influence of amount of data in statistical speech synthesis and the effect of sharing data in bilingual voices.

Keywords: META-NET, speech synthesis, multilingual synthetic voices, Festival TTS, unit-selection, statistical synthesis

1 Introduction

META-NET, a Network of Excellence consisting of 54 research centres from 33 countries, is dedicated to building the technological foundations of a multilingual European information society (MET, 2012). One of the parts of META-NET is META-SHARE, a sustainable network of repositories of language data, tools and related services, where data and tools will be both open and with restricted access rights.

Text-to-speech (TTS) is the artificial production of speech. This is an active area of research with still many challenges to improve quality, naturalness, flexibility and expressiveness of the synthetic voices. However, this is also a mature technology with many available products including TTS. Making synthetic voices widely available, either as a linguistic resource or as a service, will extend the use of TTS in many applications, in particular in the WEB and in the Mobile environments.

Recently, UPC produced and distributed synthetic voices for Catalan (Bonafonte et al., 2008; Bonafonte, 2007) as open-source resources. The voices have been included in several open-source distributions and applications. In this paper we present our work to produce and distribute Spanish and bilingual (Spanish/English) synthetic voices in the META-NET framework.

In Section 2 we will describe some resources which were produced in the TC-STAR (TCS, 2004 2007) project and which have been used to build the voices. Section 3 describes the voice generation process. In order to compare speech synthesis technologies and to document the distributed voices, we have evaluated the quality of the synthetic voices. The results are shown in Section 4. Finally, Section 5 summarizes the paper and presents some conclusions.

2 Linguistic Resources

In the previous project TC-STAR (TCS, 2004 2007), speech databases were produced for supporting the research on speech synthesis, voice conversion and synthesis of expressive speech in speech-to-speech translation. These databases were built for different languages (English, Mandarin and Spanish) following common specifications (Bonafonte et al., 2006). UPC produced Spanish and Spanish/English databases which have been included in META-NET. Table 1 shows the main characteristics of these databases.

Corpus Content	Lang.	Spk.	Rec.Time
TTS Baseline Voices: Parliamentary transcribed speech, novels, news, special sentences and words, etc.	Spanish	F0	11h30m
		M0	10h46m
Intralingual voice conversion: Short sentences in mimicking style	Spanish	F1	15m
		F2	16m
		M1	16m
		M2	16m
	English	F1	14m
		F2	14m
		M1	13m
		M2	15m
Crosslingual voice conversion: Parallel corpus in reading style	Spanish	F1	1h07m
	English		1h05m
	Spanish	F2	1h24m
	English		1h16m
	Spanish	M1	1h01m
	English		59m
	Spanish	M2	1h08m
	English		58m
Expressive speech Parallel parlamentary corpus read in expressive style	Spanish	F1	1h02m
	English		1h
	Spanish	F2	1h03m
	English		1h
	Spanish	M1	1h
	English		1h
	Spanish	M2	58m
	English		53m

Table 1: TC-STAR Spanish databases.

Two professional speakers (female F0 and male M0) recorded more than 10 hours of speech. The database was designed to produce high-quality synthetic voices using concatenation of selected units. The corpus was designed to get large variability, including many phonetic and prosodic contexts. Therefore, this resources can also be used to build voices with other technologies, as parametric statistical synthesis (Zen et al., 2009). The speakers were selected from 10 professional candidates (5M+5F) in base of the pleasantness of their voice, quality of the signal after speech manipulation and quality of synthetic sentences built from the *selection* data. The database was annotated with orthographic, phonetic, prosodic and pitch tiers.

Furthermore, four bilingual speakers (F1, F2, M1, M2) recorded several corpus to support the research on intralingual and interlingual voice conversion. First, the speakers recorded a set of sentences (around 15 min.) in a mimic style: the speaker listened to a *voice template* before repeating the sentence with similar speed and intonation. In this way, it was possible to investigate on voice conversion techniques focusing on the segmental aspects.

The speakers also recorded a parallel corpus which was original selected in English and then translated into Spanish. The speakers first recorded the English sentences and then the Spanish translation. The goal was to support the research on cross-lingual voice conversion.

Finally, to produce more expressive data, the bilingual speakers recorded a parallel corpus which was selected from the transcriptions of the European Parliament in different situations. The speakers listened to the parliamentary and they read a short paragraph, first in Spanish and then in English in the same speaking style than the parliamentary, *acting* as parliamentarians.

All these resources were produced in a recording studio with semi-professional equipment: three channels were synchronously recorded (membrane microphone, close-talk microphone, glottograph), with 96kHz as the sampling frequency and 24 bits per sample resolution.

These resources are distributed through ELDA and META-NET so that different labs can build high-quality voices in Spanish and also bilingual voices.

3 Building voices

The second resources to be included in META-NET are synthetic voices. A TTS (text-to-speech) system is an automatic system that can read a written text using what is known as synthetic voices. Several technologies have been proposed, being concatenation of selected units (Hunt and Black, 1996) and statistical/parametric synthesis (Tokuda et al., 2000), the dominant techniques.

In order to make the voices widely usable, we have created voices for Festival (Black et al., 1996 2009), a well-known open-source speech synthesizer included in many platforms.

For both *baseline* speakers, F0 and M0, synthetic voices have been created using unit-selection technology (Hunt and Black, 1996). In this technology, speech segments (*units*) are *selected* from the large database by choosing, for each phoneme, the *best unit* in the speech database, taking into account both the linguistic contexts (target cost) and the continuity between selected segments (concatenation cost).

Recently, statistical/parametric modeling has become an important technology to produce synthetic speech (Tokuda et al., 2000; Zen et al., 2009). Hidden Markov Models (HMM) are estimated for each context-dependent phoneme using the speech corpus. The context features include many phonetic and prosodic features which allow to define and estimate context-specific models. In the operative phase, these models are used to generate synthetic speech from text, i.e., from the phonetic and prosodic features, which can be easily derived from the text. One of the advantages of this technology is that it can manage data from different sources to estimate the models. Techniques as clustering models, speaker normalization and model interpolation allow to share data coming from different languages, different speakers and different styles. Furthermore, the footprint of these voices are usually several order of magnitude smaller than unit selection voices. This is an important advantage in many applications and platforms. In the project we have developed two additional voices for speakers F0 and M0 using HTS statistical/parametric models, which is also integrated into Festival.

Furthermore, the bilingual speakers F1, F2, M1, M2, produced speech in two languages (Spanish, English) and with different styles (mimic voice, reading, parliamentary style). Although the amount of data and the variability makes difficult to use it for concatenating synthesis, it can be used in the statistical framework. Four bilingual synthetic voices using statistical/parametric models have been produced. Table 2 summarizes the synthetic voices that are being included in META-NET.

Speaker	Technology	Language
F0 M0	Unit-selection based voices	Spanish
F0 M0	HMM-based voices	Spanish
F1 F2 M1 M2	HMM-based voices	Bilingual Spanish/English

Table 2: META-NET Voices

Two options have been considered.

- On one hand, for each speaker and for each language, monolingual independent voices have been produced. Therefore, the bilingual voices can be created from two sets of monolingual HMM.
- Alternatively, for each speaker, the Spanish and the English data have been jointly used to train bilingual HMM. The Spanish and the English phone sets have been mapped so that equivalent phones share the same representation in both languages. The language of the utterance is save as an additional feature of the phoneme.

In this work we have given priority to the easy distribution of the results. Festival voices have been selected as

this is a systems widely included in many open source distributions and many accessibility applications use Festival voices. We are aware that this is not the best achievable quality. For instance, for unit-selection, we are using the Festival module known as CLUNITS. This module *clusters* simplifies the search algorithm: instead of of considering all the units and computing their target cost, all the units in a pre-computed cluster are the candidate units: the decision of the *best* units is based on the concatenation cost. With respect to statistical synthesis, we also use the *HTS* module included in Festival, which is not the best voice which can be obtained even with HTS. For instance, one limitation of this module is the speech signal model, which is a basic LPC vocoder. Recently, several speech models have been proposed (see for instance (Kawahara, 2006; Erro et al., 2011)) that increase significantly the quality of statistical/parametric voices.

As it was already explained we built two models of CLUNIT voices for one male spanish speaker and one female Spanish speaker. In order to build these models we used the scripts of the Festival-TTS Synthesizer. The phonetic transcription was based on rules because for the Spanish language this is the easiest way, except for some special words which are considered exceptions. The phoneset used was taken from the SAMPA-computer readable phonetic alphabet (Wells and others, 1997). The segmentation of the speech database into phonemes was done using Ramses, our in-house speech recognition system (Mariño et al., 2000) as proposed in (Adell et al., 2005). The systems computes the best forced alignment with some flexibility to select the best pronunciation (in case the lexicon includes several ones) or to detect pauses. Afterwards, we built the trees of to cluster speech units using the CLUNITS scripts of Festival. The voices are ready for being used in the Festival-TTS platform. The weight of the voices is around 200MB, which is dominated by the raw speech data

Moreover, using the same data, we built two voices using *hts* (HTS, 2001). From previous segmentation (using Festival front-end), the required context-dependent lab files are created so that the statistical models can be estimated. The phoneset and the phonetic transcription rules is the same than the one defined for the CLUNITS voices. Context-dependent HMM are estimated to model log F0, mel-cepstrum, global variance and the state duration. These models can also be included in Festival-TTS System.

In addition, eight statistical monolingual-built models are created using either the Spanish data or the English data of speakers M1, M2, F1, F2. The procedure is the same than the one described above for M0 and F0. For the English voices, we also used SAMPA (Wells and others, 1997). The OALD dictionary provided in Festival was used.

Finally, we also developed bilingual-built HMM-voices from the same databases used before. For each speaker, all the English and Spanish data is used. The *label* files include an additional feature to indicate the language in which the phoneme was uttered. One of the additional questions provided to the clustering algorithm is about the language. In this way, for each phoneme in a given context, the clustering process decides if it is better represented by language-dependent or language-independent model. The analysis of the clustering trees show that the question concerning the language is not always the first one being used. Other additional questions put together properties shared by some phones of both languages (common and non-common phones). Once the bilingual HMM are trained, they can be used to generate synthetic speech either in English and in Spanish, including the language feature in the context-dependent labels.

4 Assessment of built voices

Last section described the different voices which have been built. Before releasing the voices, three evaluation tests have been performed to select which option is released and to provide the user information about the performance of the voices. Three tests have been done using a web interface: the judges are asked to listen to synthetic sentences and rate them. A text box has been included so that the participants can comment about the voices or the test. The number of people which participated in each test was ≈ 45.

CLUNITS vs. HTS

In principle, the voices with the best quality should be trained with the baseline speakers F0 (female) and M0 (male). These speakers were selected specifically for building unit-selection synthetic voices. Furthermore, the amount of data is significantly large.

The first test compares the unit-selection voices with the statistical/parametric voices. The state of the art unit-selection systems are still the best ones, but the difference is reduced every year. However, let's note that we are using open-source technology (CLUNITS, from Festival, and *LPC vocoder* in HTS) so we expect to have worse results than state-of-the-art systems.

The voices generated using these two technologies are clearly identifiable. Therefore, instead of presenting several sentences and ask the participants to provide several correlated judgments, a long audio file has been generated using each technology. The participants can listen as much time as they need before selecting which of the audio files they prefer. They are not informed about which audio file correspond to which technology. The results are presented in Table 3.

Options	# times selected	
	M0	F0
CLUNITS voice is much better	22	13
CLUNITS voice is slightly better	9	9
Both voices are similar	1	2
HTS voice is slightly better.	4	6
HTS is much better.	7	13

Table 3: Preference test showing number of judges that selected either CLUNITS or HTS for the male and the female baseline voices.

As we can see, for the male voice (M0), 72% of the participants prefer the CLUNITS voice. However, this is not the case for the female voice, where there is not a clear preference.

Many participants commented that the voices which correspond to the CLUNITS voices, have in fact better quality, but there are many errors in the stress and pauses. New voices are being developed to avoid the stress issue. We expect that the quality of the released CLUNIT voices will significantly improve in the release version.

Monolingual vs. Bilingual data

The second preference test analyzes if it is worth to use all the available data (Spanish and English) or it is better to use only monolingual data for training the models. Both English and Spanish voices are distributed through META-NET. However, in this test we will focus only on the Spanish voices.

For each of the four bilingual speakers (M1, M2, F1, F2) the subjects are presented with three pairs of sentences. Each pair is made of one utterances generated with the voice trained from monolingual data and the same sentence with the voice trained from bilingual data. The utterances are presented in randomized order. The number of judgments for each speaker are $3 \times 49 = 147$. The listeners have to indicate their preference from the following options:

- The first one is much better.
- The first one is slightly better.
- Both audio files are of similar quality.
- The second one is slightly better.
- The second one is much better.

The results show that there is not a clear preference between both approaches. In half of the cases, the *similar* quality was selected and less than 10% selected the *much-better* options. Furthermore, in the optional comments, many participants expressed that both voices were very similar and that is was very difficult to make a decision. We conclude that the followed methodology does not exploit the additional data. However, the bilingual models are smaller than adding the monolingual ones. Therefore, there is a saving in the required memory.

Quality of the Voices

The objective of the last test is to rate the voices both in absolute terms and relative to the other voices. In the first part of the test, the participants evaluate the male voices and, in the second part, the female voices. For each speaker, only one long audio file is presented and there is only one global question about the pleasantness: *rate the quality of the voice: how much you like the voice*. The revised MOS scale (R-MOS) is used: seven ratio buttons are presented and only the extreme buttons are labeled with *very bad quality* in the left and *very good quality* on the right (see Figure 1). The instructions ask the participant to make a join evaluation of the three speaker voices, listening the three audio signals as much as it is needed. The participant selection is mapped to 1–7 scale. For baseline voices, M0 and F0, the CLUNIT voices are used, as this technology produced the best results. For the other voices, the models trained with bilingual data have been used.

Figure 1: WEB interface for the 3rd test: quality of voices

Figures 2 and 3 show the pleasantness of these models using the MOS-R scale. These plots can be used to select the released voices.

We can see that the score for the CLUNIT-voices is good for the female voice but poor for the male voice. We are investigating if the reason is the stress problem (see discussion above).

With respect to the HTS models the score range from 3 to 4.5. While this is an acceptable quality, new voices should be developed as soon as advances on statistical/parametric synthesis are released under open-source license.

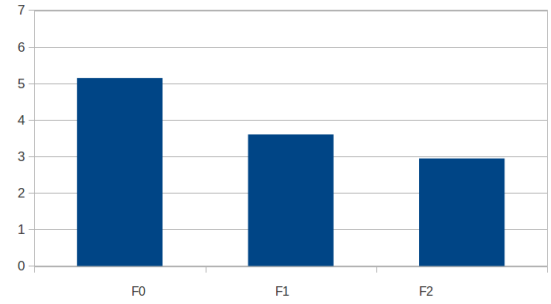

Figure 2: *Pleasantness* of the female voices in a seven-points scale.

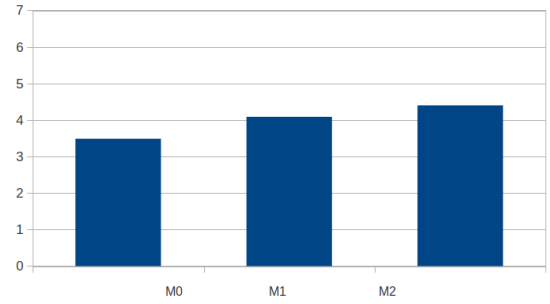

Figure 3: *Pleasantness* of the male voices in a seven-points scale.

5 Summary

In this paper we have presented the speech data that has been included in the META-SHARE platform for building synthetics voices. The data includes more than 10h of speech produced by each of two professional Spanish speakers. The database includes the needed annotation for building synthetic voices (orthographic, phonetic, prosody, segmentation, pitch labeling). Furthermore, data from 4

bilingual speakers (Spanish/English) has been delivered. This data can be easily used to estimate statistical models for parametric speech synthesis.
The paper also explain the process followed to create CLUNIT and HTS voices to be included in Festival. Although this system still does not include all the state-of-the-art technologies, it is distributed as open-source and Festival voices are easily installed and very useful.
The paper present an evaluation of the built voices which show that:

- For the baseline voices (M0 and F0) CLUNIT voices provide better results, compared with HTS voices. This is clearly expressed in the comments of the participants in the evaluation.
- The use of bilingual data did not improve the quality significantly, but it reduces the footprint of multilingual systems, as the same models can be used for Spanish and English.
- Although using more data can produce better synthesis, the pleasantness of the voice depends also of the original speaker and of some implementation issues. In particular, the bilingual male speaker seems to be preferred to the baseline one. A new evaluation should be done with new CLUNIT voices which have solve the stress problem detected.

6 Acknowledgment

The research leading to these results has been partially funded by the European Commission through the contract the METANET$_4$U(grant agreement no. 270893).

7 References

Jordi Adell, Antonio Bonafonte, Jon Ander Gómez, and María José Castro. 2005. Comparative study of automatic phone segmentation methods for TTS. In *Proc. of ICASSP*, Philadelphia, PA, USA, March.

A. W. Black, P. Taylor, and R. Caley. 1996–2009. The festival speech synthesis system.

Antonio Bonafonte, Harald Höge, Imre Kiss, Asunción Moreno, Ute Ziegenhain, Henk van den Heuvel, Horst-Udo Hain, Xia S. Wang, and Marie-Neige Garcia. 2006. TC-STAR: Specifications of language resources and evaluation for speech synthesis. In *Proc. of LREC Conf.*, pages 311–314, Genoa, Italy, May.

Antonio Bonafonte, Jordi Adell, Ignasi Esquerra, Silvia Gallego, Asunción Moreno, and Javier Pérez. 2008. Corpus and voices for catalan speech synthesis. In *Proc. of LREC Conf.*, pages 3325–3329, Marrakech, Morocco, May.

Antonio Bonafonte. 2007. FestCat: Catalan corpus and voices for speech synthesis. `http://www.talp.cat/festcat`.

D. Erro, I. Sainz, E. Navas, and I. Hernáez. 2011. Improved hnm-based vocoder for statistical synthesizers. In *Proc. of INTERSPEECH*, pages 1809–1812, Florence, Italy, September.

2001. Hmm-based speech synthesis system HTS. `http://hts.sp.nitech.ac.jp/`.

A. Hunt and A. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. of ICASSP*, volume 1, pages 373–376. Atlanta, Georgia.

H. Kawahara. 2006. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, 27(6):349–353.

José B Mariño, Albino Nogueiras, Pau Pachès-Leal, and Antonio Bonafonte. 2000. The demiphone: An efficient contextual subword unit for continuous speech recognition. *Speech Communication*, 32(3):187–197, October.

2012. META: A network of excellence forging the multilingual europe technology alliance. `http://www.meta-net.eu/`.

2004–2007. TCSTAR: Technology and corpora for speech to speech translation. `http://www.tcstar.org`.

K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. of ICASSP*, volume 3, pages 1315–1318, Istanbul, Turkey, June. IEEE.

J.C. Wells et al. 1997. SAMPA computer readable phonetic alphabet. `http://www.phon.ucl.uk/home/sampa.htm`.

H. Zen, K. Tokuda, and A.W. Black. 2009. Review: Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064.