Aholab Speech Synthesizer for Albayzin 2012 Speech Synthesis Evaluation

Iñaki Sainz, Daniel Erro, Eva Navas, Inma Hernáez, Jon Sánchez, Ibon Saratxaga.

Aholab Signal Processing Lab - University of the Basque Country, Bilbao, Spain
{inaki, derro, eva, inma, ion, ibon}@aholab.ehu.es

Abstract. This paper describes the Text-to-Speech (TTS) system presented by Aholab-EHU/UPV in the Albayzin 2012 Speech Synthesis (SP) evaluation campaign. Due to the characteristics of the provided corpus (multiple emotions and little data per emotion) a statistical synthesis approach was selected, training a single model with all the speech available. Various Pitch Detection Algorithms (PDAs) were employed in order to reduce gross errors and a more robust duration prediction was achieved by means of combining several machine learning methods. AhoCoder, the vocoder developed by Aholab Laboratory, was used to parameterize and reconstruct the speech signal with high quality.

Keywords: speech synthesis, statistical synthesis, emotions

1 Introduction

The Albayzin SS evaluation compares the performance of different TTS systems built with a common Spanish speech database. This year the database includes as a novelty a parallel corpus with four emotions (happiness, sadness, surprise and anger) plus neutral style. Once the training data is released, participants have several weeks to build the voice. Then, each group is asked to synthesize several hundred test texts that will be evaluated to determine the quality of the synthetic voices in terms of: naturalness, similarity to the original speaker, identification of the intended emotion and the perceived emotional intensity.

AhoTTS [1] is the synthesis platform for commercial and research purposes that Aholab Laboratory has been developing since 1995. It has a modular architecture, and written in C/C++ it is fully functional in both UNIX and Windows operating systems. Up to this date, synthetic voices for Basque, Spanish and English languages have been created. Contrary to last edition in which we submitted two systems (a statistical one and a hybrid system), this year we only present a statistical one. The reasons for this approach will be explained later on.

This paper is organized as follows. First, we describe the characteristics of the system. In Section 3 the voice building process is explained. The evaluation results are presented and discussed in Section 4. And finally, some conclusions are drawn in Section 5.

2 System Overview

This year Aholab presents a statistical TTS based on HTS [2]. As HTS does not perform any kind of linguistic analysis, the output of the first module of AhoTTS had to be translated into proper labels containing phonetic and linguistic information.

The reasons for not using the concatenative hybrid system [3] that yielded good results in Albayzin2010 [4] are several. The hybrid system showed good naturalness and consistency in part due to the characteristics of the recordings provided. (e.g. the low pitch variability of the male speaker and the availability of more than 2 hours of recordings in the same style). Thanks to those characteristics almost no pitch modification was necessary in the selected sequence of natural units, preserving, that way, their naturalness and reducing the distortion due to incorrect pitch marking. In the Albayzin 2012 data, those characteristics are missing because of the larger pitch variance of emotions as happiness and surprise. Besides, the amount of data per emotion is smaller than in the past edition, thus, increasing the probability of not finding proper natural units in the corpus and introducing distortions by having to modify their prosody. The possibility of looking for the optimal units in all the corpus available (i.e. including all the emotions) was also discarded due to some past experiments with similar data, in which the concatenation of units from different recording styles yielded unnatural sound most of the time.

Therefore, statistical synthesis was the chosen method by Aholab laboratory for Albayzin 2012 SS evaluation campaign. On the one hand it usually produces a more robust synthetic voice with little training data thanks to the good generalization of the models for unseen contexts. On the other, it offers greater flexibility [5] to combine all the available data from different styles, as long as it is properly labeled.

Our system was built using all the emotional data at once by simple adding a sentence level emotional label which indicated the emotion portrayed by the speaker during the recording session. With that procedure the statistical training algorithm was capable of combining data from different styles when they were similar enough, and otherwise modeling emotion specific characteristics.

2.1 Linguistic Processing

This first module performs several language dependent tasks. Text normalization and grapheme to phoneme conversion are conducted by means of rules, whereas POS tagging uses a specific lexicon and some simple disambiguation rules. The following features have been encoded into the context labels used by HTS. In short, there are the same features used in Albayzin 2010 removing the Intonation Break information and adding the emotion/style one:

- **Phoneme level:**
  - SAMPA label of the current phoneme.
  - Labels of 2 phonemes to the right and 2 phonemes to the left.
  - Position of the current phoneme in the current syllable (from the beginning and from the end).

- **Syllable level:**
  - Number of phonemes in current, previous and next syllables.
  - Accent in current, previous and next syllables.
  - Stress in current, previous and next syllables.
  - Position of the current syllable in the current word (from the beginning and from the end).
  - Position of the current syllable in the current accent group.
  - Position of the current syllable in the current sentence.
  - Position of the current syllable after the previous pause and before the next pause.
- **Word level:**
  - Simplified part-of-speech tag of the current, previous and next words (content/function).
  - Number of syllables of the current, previous and next words.
  - Position of the current word in the sentence (from the beginning and from the end).
  - Position of the current word after the previous pause and before the next pause.
- **Accent level:**
  - Type of current, previous and next accent groups, according to the accent position.
  - Number of syllables in current, previous and next accent groups.
  - Position of the current accent group in the sentence (from the beginning and from the end).
  - Position of the current accent group after the previous pause and before the next pause.
- **Pause context level:**
  - Type of previous and next pauses.
  - Number of pauses to the right and to the left.
- **Sentence level:**
  - Type of sentence.
  - Number of phonemes.
  - Number of syllables.
  - Number of words.
  - Number of accent groups.
  - Number of pauses.
  - Type of emotion/style.

2.2 Prosody Prediction

The intonation prediction is performed by the robust Multi-space Distribution MSD HMM [6], whereas the duration is predicted externally by a combination of CART and Random Forests (RF) [7] machine learning techniques. Durational data was divided into three broad classes (vowels, voiced consonants and unvoiced consonants) and optimal variables, such as the minimum occupancy per leaf or the number of trees (in RF), were determined by means of a ten-fold CV procedure. Table 1

shows the best results (i.e. using RF for vowels and CART for the other two classes) measured by the Pearson Correlation Coefficient.

Phoneme Class	Correlation (Train)	Correlation (Test)
Vowels	0.922	0.814
Voiced Consonants	0.833	0.784
Unvoiced Consonants	0.843	0.780

Table 1. Duration Prediction with 10-fold CV

We did not perform any kind of break insertion at synthesis time. So, only ortographic punctuation marks were considered as pauses.

2.3 Waveform Generation

AhoCoder [8], a Harmonic plus Noise Model (HNM) based vocoder was used in order to generate the synthetic speech from the estimated parameter sequence: spectrum, pitch and maximum voiced frequency (MVF).

3 Voice Building

Organizers provided a Spanish Emotional Voices (SEV) database recorded by GTH [9]. It contains data from the professional male speaker (Joaquín), simulating four full-blown emotions (happiness, sadness, anger and surprise) and a neutral speaking style, in a studio recording environment. The corpus for development contains more than three hours of acted-emotions recordings (approximately 40 minutes per style). In addition to the audio data and texts, ElectroGlottoGraph (EGG) waveform files were also available.

The training process of the statistical parametric voice is automatically done, once context labels and segmentation marks are ready and proper questions to build the trees are set.

3.1 Segmentation

Although the organizers provided segmentation labels, we decided to segment the whole corpus again with HTK toolkit [10] using the phoneme sequence provided by our own grapheme to phoneme converter. First, monophone models were trained from a plain start. Then, tied-state triphone models were trained and new labels obtained by means of forced alignment. During this first pass, the insertion of short pauses (SPs) was allowed at word boundaries. As an automatic post-processing before the second pass, pauses smaller than a threshold were eliminated and phoneme durational outliers were analyzed to decide whether to insert a pause next to them or not. Models were retrained after these changes and a second forced alignment was performed. This time no SPs were allowed. Finally, pause boundaries were automatically refined with a

simple processing based on phone duration and energy threshold. No manual revision of the segmentation labels was done.

3.2 Feature Extraction

All the language related features were extracted from our linguistic processing module. The extraction of the acoustic features consists of several steps. First, all the signals (speech and EGG) were down-sampled to 16kHz. Then, power normalization was performed by measuring the mean power in the middle of the vowels for all the sentences, and then normalizing each interpause interval. Afterwards, pitch contours were detected combining three different methods in order to avoid gross errors (our own PDA [11], get_f0 from Snack Toolkit and Praat). As far as the HTS training is concerned, the following parameters were extracted: f0 + 40 MFCCs + MVF.

Informal listening tests were done comparing the synthesis results for different pitch estimation inputs, and the best performance was obtained by extracting the pitch from the EGG signal with the three different PDAs and refining the combined pitch during the harmonic analysis of AhoCoder. The only exception to that rule was the surprise emotion for which the EGG recording was quite noisy. Therefore, the pitch was extracted from the speech signal for this emotion.

3.3 Impact of 'type of emotion' Label

As mentioned before, in order to use all the data to train the statistical models a new label was added indicating the type of emotion at sentence level. That way, we tried to overcome the little data per emotion available. To discover the importance of this new label we analyze the spectrum and pitch trees for all their five states. As far as pitch is concerned, the first appearance of the label in a question is before the third level in any of the trees. That shows that the tree leafs are rapidly clustered depending on their emotional content, especially for surprise and sadness emotions (which are the ones with highest and lowest variance respectively). As far as the spectrum is concerned, similar results are obtained. The question 'type of emotion' appears before the third level of the trees, and sadness and surprise are the most discriminating emotions (both had a distinctive voice quality).

Informal listening tests indicated that using all the available emotional data to build a single model produced a more consistent sound than building separate models per emotion. Being the sentences uttered by the same speaker, recorded in the same conditions and having each emotion correctly identified, it seems that the statistical modeling is able to make more robust models by clustering the similar frames of speech coming from different emotions.

4 Evaluation Results

Participants were asked to synthesize more than one hundred sentences for each emotion plus neutral style. Listeners performed 4 evaluation tasks: (i) Mean Opinion

Score (MOS) for the overall naturalness or quality, (ii) MOS for similarity to the original speaker, (iii) identification ability for the intended emotion and (iv) MOS for the perceived emotional intensity or strength.

All these four average metrics were normalized (into a range from 0 to 1) and a single measure called Emotional Performance measure (EP-measure) was defined in order to determine the best system. EP-measure is a combination of all the four metrics into one final score per system.

In this edition, 4 systems took part in the evaluation campaign. Each system was identified by a letter from B to E, being A reserved for the natural speech. Our synthetic system was identified by letter B and got the best EP-measure for neutral style, and the best average normalized speaker similarity. The results for each task are analyzed in the following subsections.

4.1 Naturalness or Speech Quality

In this task the overall quality of the speech was evaluated in a 5-point scale ranging from 1 ("very bad") to 5 ("very good)". This scale was then linearly transformed to a [0.0 - 1.0] range. Our system got an average 0.44 score, ranging from 0.49 for anger to 0.39 for sadness. The slightly worse results for sadness could be explained by the difficulty to properly extract the pitch and the voiceness for this emotion due to the special voice characteristic that it usually has.

4.2 Similarity to the Original Speaker

Listeners had to rate how similar synthetic or natural speech was when compared to the neutral voice of the speaker, in a 5-point scale ranging from 1 ("Sounds like a totally different person") to 5 ("Sounds like exactly the same person"). Our system got a median value of 3 points for all the emotions but for sadness, for which 2 points were scored. As far as the normalized speaker similarity is concerned, our system got the best results with an average of 0.46, ranging from 0.39 for sadness to 0.52 for neutral and happiness. Being our system a statistical one, we are a bit surprise by the fact that we got the best average results. That might mean that the rest of the systems are not concatenative ones and that they might have adapted their voices from an average model built with external data.

4.3 Intended Emotion

Listeners had to identify the intended emotion from a limited set of emotional categories: happiness, anger, surprise, sadness, neutral or another. Table 2 shows the confusion matrix for system B. All the 4 emotions plus neutral style were identified with a rate much greater than chance (0.166, taking into account "other"). That rate ranges from 0.35 for surprise to a far high 0.83 for neutral style. Sadness has also a good identification rate in spite of getting the worse results in the previous tasks. Besides, most emotions are confused with neutral style. The average emotion identification rate is 0.53.

Intended Emotions	Identified Emotions					
	Happiness	*Anger*	*Surprise*	*Sadness*	*Neutral*	*Other*
Happiness	**0.38**	0.12	0.19	0.06	0.21	0.04
Anger	0.02	**0.48**	0.06	0.10	0.34	-
Surprise	0.16	0.06	**0.35**	0.18	0.18	0.08
Sadness	0.04	0.02	0.02	**0.56**	0.32	0.04
Neutral	0.04	0.02	-	0.10	**0.83**	0.02

Table 2. Confusion matrix for our system

4.4 Perceived Emotional Intensity

Listener had to assess the emotional strength or intensity in a 5-point scale from "very weak" to "very strong". The average strength score of our system is 0.41, ranging from 0.33 for neutral to 0.47 for happiness and sadness. It is somehow unexpected that being neutral the "emotion" with highest identification rate, it gets the worse emotional intensity. Maybe listeners were confused when having to evaluate the emotional intensity of a neutral or non emotional speech.

5 Conclusions

This has been our third participation in the Albayzin SS evaluation campaign. We built a statistical synthetic voice, using all the available emotional data and including a context label in order to identify each emotion at sentence level.

4 systems took part in the evaluation and ours got the best EP-measure for neutral style the best average normalized speaker similarity. All the emotions were identified far above chance levels and we could state that our approach training a single voice with all the available emotional data has succeeded. As in previous editions, there is still a big gap between the best synthetic system and natural speech, but that gap seems to be even larger when dealing with emotional data.

6 Acknowledgements

The authors would like to thank the organizers of Albayzin SS 2012 and the developers of all the tools employed during the voice building process.

This work has been partially supported by UPV/EHU (Unidad de Formación e Investigación UFI11/30), the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-02) and The Basque Government (Saiotek Project, PE11UN081).

7 References

1. Hernaez, I., Navas, E., Murugarren, J.L., and Etxebarria, B. Description of the AhoTTS conversion system for the Basque language. In: Proceedings of 4th ISCA Tutorial and Research Workshop on Speech Synthesis. (2001)
2. "HMM-based Speech Synthesis System (HTS)," http://hts.sp.nitech.ac.jp/
3. Sainz, I., Erro, D., Navas, E., Hernáez, I., Sánchez, J., Saratxaga, I., Odriozola, I., and Luengo, I. Aholab Speech Synthesizers for Albayzin2010. In: Fala2010. pp. 343--348 (2010)
4. Méndez, F., Docío, L., Arza, M., and Campillo, F. The Albayzín 2010 Text-to-Speech Evaluation. In: Fala2010. pp. 317--340 (2010)
5. Zen, H., Tokuda, K., and Black, A.W. Statistical parametric speech synthesis. In: Speech Communication. vol. 51. no. 11. pp. 1039--1064 (2009)
6. Tokuda, K., Masuko, T., Miyazaki, T. Multi-space probability distribution HMM. In: IEICE Trans. Inf. & Syst. vol. E85-D. no. 3. pp. 455--464 (2002)
7. Breiman, L. Random forests. In: Machine learning. vol. 25. no. 2. pp. 5--32 (2001)
8. Erro, D., Sainz, I., Navas, E., and Hernáez, I. HNM-Based MFCC+f0 Extractor Applied to Statistical Speech Synthesis. In: ICASSP 2011. pp. 4728--4731 (2011)
9. Barra-Chicote, R., Montero, J.M., Macias-Guarasa, J., Lutfi, S., Lucas, J.M., Fernández, F., D'Haro, L.F., San-Segundo, R., Ferreiros, J. , Córdoba, R., Pardo, J.M. Spanish Expressive Voices: corpus for emotion research in Spanish. In: Proceedings of the 6th conference of Language Resources & Evaluation (Workshop on Corpora for Research on Emotion and Affect). pp 60--70. Marrakech, Morocco (2008)
10. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Olla-son, D., Povey, D., Valtchev, V., and Woodland, P.C. The HTK Book. version 3.4 (2006)
11. Luengo, I., Saratxaga, I., Navas, E., Hernáez, I., Sánchez, J., and Sainz, I. Evaluation of Pitch Detection Algorithms Under Real Conditions. In: IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP'07. pp 1057--1060 (2007)