# Versatile Speech Databases for High Quality Synthesis for Basque

**I. Sainz, D. Erro, E. Navas, I. Hernáez, J. Sanchez, I. Saratxaga, I. Odriozola**
Aholab – Dep. of Electronics and Telecommunications. Faculty of Engineering
University of the Basque Country, Urkijo zum z/g 48013 Bilbo
E-mail: {inaki, derro, eva, inma, ion, ibon, igor}@aholab.ehu.es

This paper presents three new speech databases for standard Basque. They are designed primarily for corpus-based synthesis but each database has its specific purpose: 1) AhoSyn: high quality speech synthesis (recorded also in Spanish), 2) AhoSpeakers: voice conversion and 3) AhoEmo3: emotional speech synthesis. The whole corpus design and the recording process are described with detail. Once the databases were collected all the data was automatically labelled and annotated. Then, an HMM-based TTS voice was built and subjectively evaluated. The results of the evaluation are pretty satisfactory: 3.70 MOS for Basque and 3.44 for Spanish. Therefore, the evaluation assesses the quality of this new speech resource and the validity of the automated processing presented.

**Keywords:** Speech Corpus, Speech Synthesis, Evaluation

## 1. Introduction

The most successful TTS (text-to-speech) systems nowadays are the corpus-based ones (i.e. unit selection and statistical parametric). In unit selection concatenative systems (Hunt & Black, 1996) the most appropriate natural units are selected from a speech database and joined together trying to reduce the concatenation artifacts. In statistical parametric systems average models are trained from acoustically similar natural units, building decision trees with linguistic features. While the concatenative approach offers a higher naturalness (especially in limited domains), the statistical one provides more stability and flexibility to create new voices through adaptation or interpolation techniques (Zen, Tokuda, & Black, 2009). And even though the size of the database is not that important for statistical systems, both technologies benefit from a large size phonetically rich corpus in the generation of high quality synthetic speech. The development of such a corpus is especially important for languages with limited resources as it is the case of the Basque language (less than 700.000 speakers). In fact, there was already a TTS database for Basque (Saratxaga, Navas, Hernaez, & Luengo, 2006), but its small size did not permit the construction of high quality prosodic and acoustic modules. In Concatenative TTS, if all the remaining aspects are kept unchanged (e.g. Voice quality within sessions) the broader the phonetic coverage, the better the performance of the system is. As far as HMM-based TTSs are concerned, though they provide a sufficient quality even for small databases, a larger corpus would certainly yield more accurate models. Besides, once you have good models for a voice, adaptation techniques would allow building new ones from already existing o newly recorded small databases of less than 100 sentences.

In this paper, the recording and annotation process of three new databases for Basque language is detailed. In section 2 we describe the specifications of the text corpus and how the recording sessions were organized. Section 3 is focused on the automatic annotation process of each database. In section 4 a subjective evaluation of an HMM-based TTS voice is run in order to assess the quality of the new speech resource. Finally, some conclusions are drawn in section 5.

## 2. Corpus Building

The corpus building process involves several steps that must be approached carefully in the goal of achieving a high quality speech database. First, the text corpus has to be designed taking into account the possible purposes of the TTS. Then, appropriate speakers must be chosen. And finally the recording must take place in proper conditions.

### 2.1 Corpus Design

Table 1 resumes the characteristics of each of the databases recorded that, together with the unrestricted domain requirement, composed our initial specifications to design an appropriate corpus. AhoSyn has been designed for high quality speech synthesis and includes one female voice and one male voice. AhoSpeakers will be used for voice conversion and includes 3 female speakers and 4 male speakers. The purpose of AhoEmo3 is emotional speech synthesis and includes speech from one male and one female speaker. The only overlapping that exists among speakers and databases is the following one: The neutral part of AhoEmo3 is also included in AhoSpeakers database.

| **Database** | *AhoSyn* | *AhoSpeakers* | *AhoEmo3* |
|---|---|---|---|
| **Language** | Basque & Spanish | Basque | Basque |
| **Purpose** | HQ Synthesis | Voice Conversion | Emotional Synthesis |
| **Gender** | 1F & 1M | 3F & 4M | 1F & 1M |
| **Style** | Neutral | Neutral | Neutral + 3 emotions |
| **Size** | 6 hours | 1 hour | 1 hour |

Table 1: Main characteristics of the databases.

First, the text corpus for AhoSyn database was constructed and then, a small portion of it was selected to record the remaining databases  So, the initial step was to compile huge amounts of textual data for each of the target languages. As the domain of the TTS was supposed to be unlimited, we tried to get texts from as many sources as possible. Being Basque a minority language, this was not an easy task to achieve. In the end, more than 400MB of plain text were collected from different domains: News (23%), Literature (19%), Arts (18%), Sciences (10%) and others. A similar text compilation was accomplished for Spanish, being a far simpler task to tackle.

To clean the initial corpus some automatic steps were taken (e.g. deletion of sentences containing foreign words whose anomalous transcription could distort the phonetic analysis). AhoTTS (Hernaez, Navas, Murugarren, & Etxebarria, 2001) system was used as the transcription tool for both Basque and Spanish languages. Next, with the help of a greedy algorithm (Sesma & Moreno, 2000) a subset of sentences were selected from the huge initial text corpus. To do so, the following criterion was used: maximize the diphone coverage according to their frequency of appearance in the collected data, limiting the number of words per sentence to less than 15 (to keep the corpus easily readable). Moreover, a parallel selection was launched only for interrogative sentences due to their peculiar intonation features (they represent approximately 14% of the corpus). All extracted sentences were proofread, discarding the invalid ones (e.g. grammatically wrong) and correcting some misspellings. The correction and selection process was repeated up to five times until obtaining the corpus described in Table 2. Table 3 shows the most frequent diphones for each language.

| **Number of...** | *Spanish* | *Basque* |
|---|---|---|
| **Sentences** | 3995 | 3799 |
| **Words** | 51380 | 38544 |
| **Distinct Phonemes** | 29 | 35 |
| **Distinct Diphones** | 539 | 583 |

Table 2:Information about AhoSyn text corpus.

As stated previously, text corpus for AhoSpeakers and AhoEmo3 databases was generated from the corpus of AhoSyn, using again the same greedy algorithm to select just 500 sentences. Among the different ways of recording emotions (i.e. spontaneous, elicited and acted), the third option was the preferred one because it offers more control over the recording conditions and the phonetic balance of the content. Its main drawback is that it can produce stereotypical and full-blown emotions, which may not be convenient for real emotion recognition but that can be adequate for building TTS voices (Navas, Hernaez, & Luengo, 2006). Moreover, statistical interpolation between neutral style and full-blown emotions can lead to different grades of intensity (Tachibana, Yamagishi, Onishi, Masuko, & Kobayashi, 2004). Therefore, during the recording of AhoEmo3 exactly the same prompts were read for neutral style, happiness, sadness and anger emotions. These were the three chosen emotions because we think that out of the “big six” (Cowie & Cornelius, 2003) they tend to be the most distinct ones (e.g. the pairs “surprise – happiness” and “fear – sadness” are quite often confused, and disgust usually has the lowest recognition rate (Scherer, 2003)). Besides, those emotional styles can be useful both for storytelling and human interface purposes.

## 2.2 Speaker Selection

The quality of a TTS is highly dependent on the speaker with which the synthetic voice is built. Several efforts have been made in order to discover the desired features that a voice talent must have (Syrdal, Conkie, & Stylianou, 1998) (Coelho, Hain, Jokisch, & Braga, 2009), with no definitive conclusion yet. We made a casting among several speakers to informally evaluate their validity for the recording of AhoSyn based on the following criteria: Voice pleasantness, clear articulation, correct pronunciation of the target language and perceptual quality of their resynthesized voice for HNM (Harmonic plus Noise Models) and PSOLA techniques. Among the remaining candidates, the ones with the best acting capabilities were selected for AhoEmo3. Not surprisingly, both were dubbing actors. Finally, voice talents for AhoSpeakers were chosen according to the uniqueness of their voice, as that would offer us a wider range of action during voice conversion experiments. It must be remarked that all the voice talents were native Basque speakers for AhoSpeakers and AhoEmo3 and bilingual as far as AhoSyn database is concerned. In total, 9 speakers were recorded: 5 male and 4 female. That means that the neutral part of AhoEmo3 was also included into AhoSpeakers database.

| ***Spanish*** | | ***Basque*** | |
|---|---|---|---|
| **Diphone** | **N. of occurrences** | **Diphone** | **N. of occurrences** |
| e-n | 4162 | e-n | 4666 |
| e-s | 3951 | t-a | 3645 |
| d-e | 3584 | a-n | 3224 |
| e-l | 3317 | k-o | 3138 |
| l-a | 2731 | t-e | 3956 |
| t-e | 2766 | a-k | 2571 |
| o-s | 2766 | e-t | 2494 |
| a-l | 2731 | a-l | 2399 |
| a-n | 2584 | t-u | 2278 |

Table 3:Most common diphones in AhoSyn text corpus.

## 2.3 Recordings

Recordings were made in a semi-professional studio built inside our laboratory. It provides good sound isolation and its interior is acoustically treated to mitigate disturbing reverberations. The recording platform employed is shown in Figure 1.

A high quality audio interface was used to feed and connect all the devices. The process was controlled through a laptop located outside the isolated room so as to avoid possible fan noise or electrical interferences.

Prompts were displayed to the speaker by means of a screen connected to the laptop. Three channels were recorded at 48 kHz sampling rate and with 16 bits of resolution: Diaphragm microphone, close-talk microphone and the glottal pulse signal from the laryngograph. A pop filter was located between the speaker and the main microphone in order to reduce the airflow pressure. Each session was monitored from the outside with the help of headphones or speakers and a recording software (NannyRecord) developed by UPC (Universitat Politècnica de Catalunya). The communication with the speaker was done via an external microphone connected to the headphones the speaker was wearing. Most of the speakers also chose to receive some feedback of their own voice through these headphones. The equipment used during the recording sessions is listed in Table 4.

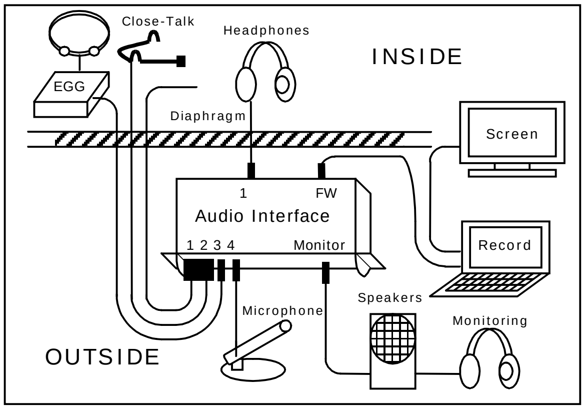

Figure 1: Recording platform.

To complete the recording of AhoSyn several sessions were necessary (while the other databases were recorded in only one session). So, in order to hinder the inter-session variability (e.g. voice quality, speed, tone, etc.) several steps were followed. The position of the microphones inside the room and the distance from the speaker to the microphones were kept almost constant during the whole recording process. Speakers were given some instructions about how to conduct their readings to reduce voice fatigue over long sessions on consecutive days. They were asked to speak effortlessly and in a volume they could sustain for a long period of time. At the beginning of each session the technician monitoring the recording would adjust the average amplitude of the input signal to a similar level to that of the last recording session. The speakers were allowed to hear a couple of sentences from past recordings so that they could maintain the rhythm and tone. In the middle of the recording, if the technician noticed that the speaker had deviated in excess from the *reference point*, new instructions were transmitted to the voice talent. Regarding the style, a natural reading style was requested for AhoSyn and AhoSpeakers databases. In AhoEmo3, a longer trial and error feedback instruction was needed for the recording of the emotional speech until the desired style was obtained. Besides, some general guidance was pointed out: try to maintain the style independent of the semantic content of the utterance and, try to be consistent with the pronunciation and careful with the prosody at breaks and sentence boundaries. If a prosodic or phonetic mistake was made during the reading of a sentence, the technician had to decide whether it was a minor or major error. Major errors involved the re-recording of the sentence, whereas minor errors were manually annotated in the prompts by the technician herself, to be processed later.

| | |
|---|---|
| **Microphones** | Neumann TLM103 (diaphragm)<br>Shure Beta54 (close-talk)<br>Philips SBC ME570 (Outside, control) |
| **Audio Interface** | RME Fireface 400 |
| **Laryngograph** | Laryngograph PCLX (LTD) |
| **Software** | NannyRecord (UPC)<br>Fireface Mixer |

Table 4: Recording equipment.

## 3. Corpus analysis

The analysis and annotation of the recorded corpora supposes a decisive process in the goal of building high quality voices. However, achieving a high accuracy usually involves a time consuming hand labeling process. We decided to combine mainly automatic labeling processes with little or none manual intervention.

First, all the waveform files were down-sampled to 16kHz and normalized in power. This normalization procedure is important to avoid excessive volume differences among voices (in case we want to build an average voice) and between different recording sessions of the same voice (AhoSyn voices). The normalization was performed per waveform in the following way: voiced portions of each signal were determined with the help of Praat (Boersma & Weenink, 2010). The mean power was then fixed to -25dV as specified in ITU-T-P.56. If the normalization led to the saturation of the signal, problematic segments were automatically detected and properly attenuated within a rectangular window. The boundaries of this window were the nearest zero crossing values outwards the problematic region itself. This simple approach reduced the excessive volume at the beginning of some utterances while preserving the natural power envelope of the sentences.

Meanwhile, initial texts selected with the greedy algorithm were corrected so they matched the sentences uttered by the speaker (this time-consuming task was only done for the Spanish recordings of the AhoSyn male voice). During the recording apart from the common reading mistakes, some consistent deviations from the standard or canonical transcription were observed. In Basque, for example, palatalization of ‘n’ and ‘l’ sounds was also done between words and some sound deletion appeared in words like (*horiek* -> *hoiek*, *them* in English), etc. Therefore, some speaker dependent transcription rules were applied. Feeding the aforementioned transcriptor with the corrected or uncorrected text files, sequences of phonemes and orthographical pauses were generated. This phoneme sequences along with the normalized signals were used to perform an automatic speaker dependent segmentation based on forced alignment.

The HTK toolkit (Young et al., 2006) was used during the segmentation process. First, tied-state triphone models were trained from a plain start, allowing the insertion of short pauses at word boundaries. Then, an automatic process was run to remove too short pauses and insert new ones. The decision of inserting new pauses was made taking into account the power envelope, the minimum duration of the pause itself as well as the duration outliers at word boundaries for each phoneme class. For example, if a duration outlier was detected at a word boundary and the amplitude around that region was below a certain threshold, a new pause was inserted. After “definitive” pauses were set, triphone models were retrained, not allowing the insertion of short pauses this time. Finally, the segmentation boundaries of the phonemes adjacent to pauses were refined by means of a simple but effective algorithm that uses power envelope and durational outliers.

As a final step, segmentation and linguistic information was automatically synchronized, removing or inserting pauses in the former files.

## 4. Evaluation of Ahosyn

In order to assess the quality of the recordings and the automatic annotation procedure, statistical parametric voices were built from scratch for the female voice of AhoSyn databases.

### 4.1 Voice building

First, speech signals were analyzed with AhoCoder, a high-quality vocoder developed in our lab (Erro, Sainz, Navas, & Hernaez, 2011). Then, proper linguistic labels were prepared (Erro et al., 2010) and the HTS system (Zen et al., 2006) was used to train HMM models.

Taking advantage of having recordings uttered by the same speaker in Basque and Spanish, and having in mind that we were using the same label structure and that both languages share most of their voiced phonemes, a single bilingual voice was built from the available material. In order to distinguish both languages an additional label was added to the statistical system at sentence level. Besides, 2 monolingual systems for Basque and Spanish were also built.

### 4.2 Evaluation design

An online evaluation campaign was organized. Listeners had to score the naturalness of synthetic sentences from the female voice of AhoSyn within a 5 point scale ranging from *1 – It sounds completely unnatural* to *5 – It sounds completely natural*. Ten texts not included in the recorded corpus were randomly selected for each language, and sentences were synthesized for two configurations: monolingual and bilingual voice. Each listener evaluated up to 20 signals: 5 out of 10 signals for each language and configuration. 18 subjects took part in the campaign, none of which had any hearing impairment. Almost all of the subjects were fluent in both languages and half of them had no experience with speech technologies. The evaluation was held in a quiet environment and all of the listeners used high quality headphones.

Before the test was conducted, some natural recordings of the speaker were presented, so as to implicitly fix the ceiling naturalness quality.

### 4.3 Evaluation results

Figure 2 shows the results of the evaluation for each method and language, including the 95% CI (Confidence Interval). Quite good results are obtained for both languages: 3.70 MOS (Mean Opinion Score) for Basque and 3.44 for Spanish. There are no significant differences in naturalness between the monolingual and bilingual approach, but the bilingual voice is 12.34% smaller than the monolingual one. It must be noted that no manual correction was performed during the automatic annotation of this voice.

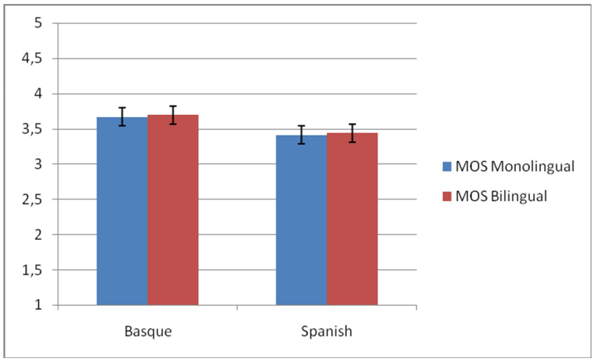

Figure 2: Subjective evaluation results.

## 5. Conclusion

Three new speech resources for Basque language have been presented. They have allowed the development of a high quality neutral TTS voice and have the potential to adapt it to a variety of new styles and voices. The design and recording of the corpus has been described in depth. Additional information about the automatic annotation process has also been included. And the subjective results of the synthetic voice built show the high quality of this new resource.

## 6. Acknowledgements

This work has been partially supported by the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-02) and the Basque Government (Saiotek Project, PE11UN081).

## 7. References

Boersma, P., & Weenink, D. (2010). Praat: doing phonetics by computer [Computer program]. Version 5.1.38.

Coelho, L., Hain, H. U., Jokisch, O., & Braga, D. (2009). Towards an Objective Voice Preference Definition for the Portuguese Language. *I Iberian SLTech 2009*, 67.

Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, *40*(1-2), pp. 5-32.

Erro, D., Sainz, I., Luengo, I., Odriozola, I., Sánchez, J., Saratxaga, I., Navas, E., et al. (2010). HMM-based Speech Synthesis in Basque Language using HTS. *Fala2010*, pp. 67-70. Vigo.

Erro, D., Sainz, I., Navas, E., & Hernaez, I. (2011). HNM-Based MFCC+f0 Extractor Applied to Statistical Speech Synthesis. *ICASSP 2011*, pp. 4728-4731.

Hernaez, I., Navas, E., Murugarren, J. L., & Etxebarria, B. (2001). Description of the AhoTTS conversion system for the Basque language. *Proceedings of 4th ISCA Tutorial and Research Workshop on Speech Synthesis*.

Hunt, A. J., & Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. *ICASSP 1996: Proceedings of the Acoustics*, *1*, pp. 373-376.

Navas, E., Hernaez, I., & Luengo, I. (2006). An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS. *IEEE Transactions on Audio, Speech and Language Processing*, *14*(4), 1117-1127.

Saratxaga, I., Navas, E., Hernaez, I., & Luengo, I. (2006). Designing and recording an emotional speech database for corpus based synthesis in Basque. *Proc. of fifth international conference on Language Resources and Evaluation (LREC)*, pp. 2126-2129.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1-2), pp. 227-256.

Sesma, A., & Moreno, A. (2000). *CorpusCrt 1.0: Diseño de corpus orales equilibrados*. [Computer Program]: http://gps-tsc.upc.es/veu/personal/sesma/CorpusCrt.php3

Syrdal, A. K., Conkie, A., & Stylianou, Y. (1998). Exploration of acoustic correlates in speaker selection for concatenative synthesis. *Fifth International Conference on Spoken Language Processing*. ISCA.

Tachibana, M., Yamagishi, J., Onishi, K., Masuko, T., & Kobayashi, T. (2004). HMM-Based Speech Synthesis with Various Speaking Styles Using Model Interpolation. *Proc. Speech Prosody*.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., et al. (2006). The HTK Book, version 3.4.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Black, A. W., Masuko, T., & Tokuda, K. (2006). The HMM-based speech synthesis system (HTS) version 2.0. *The 6th International Workshop on Speech Synthesis*, pp. 294-299.

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