EVALUATION OF A SPOKEN PHONETIC DATABASE IN BASQUE LANGUAGE

V. Guijarrubia, I. Torres, L.J. Rodríguez

Departamento de Electricidad y Electrónica. Universidad del País Vasco
Apartado 644. 48080 Bilbao. Spain
{vgga, manes, luisja}@we.lc.ehu.es

Abstract

In this paper we present the evaluation of a spoken phonetic corpus designed to train acoustic models for Speech Recognition applications in Basque Language. A complete set of acoustic-phonetic decoding experiments was carried out over the proposed database. Context dependent and independent phoneme units were used in these experiments with two different approaches to acoustic modeling, namely discrete and continuous Hidden Markov Models (HMMs). A complete set of HMMs were trained and tested with the database. Experimental results reveal that the database is large and phonetically rich enough to get great acoustic models to be integrated in Continuous Speech Recognition Systems.

1. INTRODUCTION

Basque is a minority language which is the official language, along with Spanish, for a Community of 2,5 million living in the Basque Country in the north of Spain. It is considered as one of the oldest European languages and includes some interesting and infrequent typological characteristics (Mitxelena, 1977). On the other hand, Basque language is more and more present in nowadays life: local administration, television, newspapers, etc. Thus, Universities and some local companies have increased the interest for the study of this language, not only from and academic point of view but also aimed to develop bilingual industrial applications in related with human language technologies: speech synthesis, recognition and understanding, dialog systems, automatic translation, etc.

In this framework, a phonetically balanced corpus was designed and acquired. Basque language has been poorly studied from a phonetic point of view. Thus, a large amount of data had been previously analyzed in order to classify the Basque phones and to obtain their frequency distribution. In Section 2. we fully describe the database. Preliminary experiments were carried out to define an adequate set of sub lexical units (López de Ipiña et al., 2000). This set consisted of 34 units and was established as the basic set of incontextual units to be used in future speech recognition systems working in Basque. This is the first phonetically balanced speech database recorded in Basque designed to train acoustic models for Speech recognition applications (López de Ipiña et al., 1998). Additional resources for Basque include the database designed by Telefónica I+D and our group (unpublished). It was aimed to train Telefonica's speech recognition systems and was acquired trough the telephone. Recently a Basque version of the SpeechDat is also available trough ELRA. It was also designed to develop applications working trough the telephone.

The aim of this work is to evaluate the design of the database. A complete set of acoustic-phonetic decoding experiments was carried out for this purpose. Section **??** describes the experiments carried out. Both, discrete and continuous Hidden Markov Models (Rodriguez and Torres, 2003) were used to build acoustic models. Speaker independent recognition rates were high enough to validate the design of the acoustic database. The phonetic balance is adequate and the number of samples of each phone is high enough to train continuous HMM?s.

Finally, several sets of triphones and biphones were also defined. Experimental evaluation is also described in Section 3.. Contextual dependent units achieved important error rates reductions. These results confirm the suitability of the database design also in terms of phone context distribution. The number of different contexts as well as the number of samples of more frequent contexts is high enough to allow such a reduction in phone recognition errors.

2. SPEECH DATABASE

The main objective of the corpus was to represent the full phone code used by nowadays Basque Country speakers. The phonetic balance was guaranteed in the database design. A set of 200 sentences was carefully selected to represent the phonetic distribution of the language. The sentences were written in standard Basque - i.e. batua Basque - according with Academic normative. This set was used for training purposes. For testing purposes a new set of 100 sentences was selected from nowadays basque narrative.

The selection of speakers was also achieved under a phonetic criterium. Thus a basic set of 46 speakers from the ”guipuzcoan” dialect region was selected to pronounce the sentences. The phonetic code used in this region is the richest one of the Basque Country and includes the phones used in all the dialect regions. As a consequence, the final database includes a large variety of sounds and guarantees the phonetic balance of the language. However, speakers leaving in the French Basque Country were not included in this database since their phone code is very different and include some French sounds.

Thus, the final Speech Database consisted of 300 different sentences uttered twice by 46 speakers resulting in a total of 27.600 utterances, about 200.000 words and 1.100.000 phones. It was recorded at 16 kHz in laboratory environment.

2.1. THE SET OF PHONES

The Basque language has not been broadly studied to perform speech recognition tasks. Thus, preliminary experiments were carried out to get a suitable set of sub-lexical units (López de Ipiña et al., 2000). The distribution of Basque Phones is not very different from the Spanish one. Both languages share the same vowel triangle (only five vowels). However, Basque includes larger sets of fricative and affricate sounds, some of them very particular and not present in any other European language.

Starting from these studies, the phone-like units showed in Table 1 were used to transcribe the complete database. This table shows the units in SAMPA (Speech Assessment Methods Phonetic Alphabet) notation. This notation was established in a European project framework (Esprit 2589, 1991) in order to define an ASCII symbol to represent any sound appearing in European Languages (Esprit Project 2589 (SAM), 1991). Table 1 also shows an a Basque word as an example of each unit.

3. EXPERIMENTAL EVALUATION

A complete set of acoustic-phonetic decoding experiments was carried out. The aim of these experiments was to validate the design of the database. Both, discrete and continuous Hidden Markov Models were used to build acoustic models.

3.1. EXPERIMENTAL SETTINGS

The speech database was parameterize into 12 Mel-frequency cepstral coefficients with delta and acceleration coefficients, energy and delta energy. Thus, four acoustic representations were defined. The length of the analysis window was 25 ms and the window shift 10 ms.

For the discrete models, the well-known LBG algorithm (Linde et al., 1980) was used to compute a representative set of 256 prototype vectors. The acoustic vectors in both the training and test data were then classified by assigning them the nearest of the 256 prototype vectors. Discrete HMMs with four observations code-books, one for each acoustic representation, were used as acoustic models in this case.

Each phone-like unit was modeled by a typical left to right non-skipping self-loop three state HMM, for both discrete and continuous models.

The evaluation criterion was defined as the number of correct phone-like units recognized divided by the total number of expected units (the sum of deleted (d), inserted (i), substituted (s) and correct (c) units) or phone recognition rate.

$$\%correct = \frac{c}{c+d+i+s} \quad (1)$$

The training corpus is composed of 200 phonetically balanced sentences uttered by 25 speakers, resulting in 9394 sentences and involving around 340000 training phonemes.

Type	SAMPA	Example
occlusive	[p]	**p**area
	[b]	**b**akea
	[t]	ai**t**a
	[d]	**d**abil
	[c]	An**tt**on
	[k]	**k**alea
	[g]	**g**oaz
nasal	[m]	a**m**a
	[n]	joa**n**
	[J]	ahalegi**n**a
	[N]	zere**ng**atik
affricate	[ts]	a**tz**o
	[ts_a]	i**ts**aso
	[tS]	**tx**ikia
fricative	[B]	a**b**estu
	[f]	a**f**aria
	[D]	a**d**arra
	[s]	**z**arata
	[s_a]	u**s**o
	[S]	**x**erra
	[jj]	**Y**on
	[x]	erlo**j**u
	[G]	a**g**ur
lateral	[l]	**l**orea
	[L]	muti**l**a
trill	[rr]	a**rr**azoia
	[r]	agu**r**ea
vowel	[i]	harr**i**a
	[e]	l**e**hoi
	[a]	**a**gur
	[o]	**o**rain
	[u]	ag**u**r
semivowel	[j]	ka**i**xo
	[w]	a**u**rpegi

Table 1: Phone-like units used to transcribe the database, in SAMPA notation. A total of 34 units plus silence were used. An example is given for each unit.

A small subset of this corpus, consisting of 197 sentences and 7200 phonemes, was manually segmented for model initialization purposes.
Then three testing corpus were defined:

- a speaker dependent and vocabulary independent set (SDVI): consisting of 100 new sentences from Basque narrative uttered by the 25 training speakers resulting in 2173 sentences and 87330 phonemes.
- a speaker independent and vocabulary dependent set (SIVD): consisting in the 200 training sentences uttered by 13 new speakers, not included in the training set, resulting in 5102 sentences and 88122 phonemes.
- a speaker independent and vocabulary independent set (SIVI): composed by the new 100 sentences coming from the Basque narrative uttered by the new 13 speakers, resulting in 1056 utterances, not included in the training set, and 42518 phonemes.

3.2. CONTEXT INDEPENDENT UNITS

The whole set of phone-like units in Table 1 was used in these experiments. A discrete and then a continuous HMM were trained for each unit using the training set presented above. Starting with discrete HMMs, both the Viterbi and Baum-Welch training procedures were used for these models. The initialization was made using hand-labeled phone segments and 6 iterations of a single-model Baum-Welch training algorithm (Rodriguez and Torres, 2003). These initial models were then recomputed in two independent ways: using the Viterbi algorithm and the embedded-model Baum-Welch algorithm (Rodriguez and Torres, 2003). Then three decoding experiments were carried out over SDVI, SIVD and SIVI testing sets. Table 2 shows the results obtained trough these experiments.

procedure\test set	SDVI	SIVD	SIVI
Viterbi	60.16	61.84	59.92
Baum-Welch	59.83	61.98	59.91

Table 2: Phone recognition rates obtained with the Viterbi and embedded-model Baum-Welch training procedure when using discrete HMMs. Three test sets were used: a speaker dependent and vocabulary independent set (SDVI), a speaker independent and vocabulary dependent set (SIVD) and a speaker independent and vocabulary independent set (SIVI).

Both training procedures yielded almost the same performance. On the other hand, phone recognition rates do not differ when using different test sets. Thus we can conclude that the acoustic models were not broadly adapted to the speaker and to the training corpus vocabulary. Similar experiments were carried out for Spanish language achieving better phone recognition rates for SIVI experiments (nearly 65% of recognition rates). (Rodriguez and Torres, 2003). This difference is probably due to the high number of phone-like units used in Basque: 35 units. In fact, only 25 phone-like units were required in Spanish.

In a second series of experiments continuous mixture densities HMMs were used. Due to their high computational cost, only the Viterbi training procedure was tested since this training reveals as good as Baum-Welch at a much lower cost (Rodriguez and Torres, 2003). In this case, four different numbers of gaussians per state and acoustic representation were used, namely 8, 16, 32 and 64. Models were initialized starting from the discrete models obtained through Viterbi training. These initial models were re-estimated using the Viterbi training procedure. Table 3 shows the phone recognition rates achieved trough these experiments for SIVD, SDVI and SIVI test sets.

A great improvement was achieved using continuous HMMs. The 8 gaussian HMMs gave 4-5 more absolute points than the discrete HMMs. The use of 16 gaussians yielded a remarkable, but smaller, improvement (2-3 points better than 8 gaussian HMMs). The most significant improvement was achieved with 32 gaussians (4-5 points better than 16 gaussians HMMs). These results are similar to those obtained for Spanish language trough the same experiments (Rodriguez and Torres, 2003). This is remarkable since, as was explained before, the number of units used to transcribe the database was 40% higher than the one used for Spanish transcription, so the a priori probability of making an error was also higher. These results show the quality of the acoustic database and validate its design. The phonetic balance is adequate and the number of samples of each phone is enough to train continuous HMMs.

M\test set	SDVI	SIVD	SIVI
8	65.40	66.57	64.67
16	67.69	68.48	66.52
32	71.60	73.42	73.08
64	71.18	71.33	69.25

Table 3: Phone recognition rates obtained through Viterbi training with continuous HMMs, using various mixture sizes (M = 8, 16, 32 and 64) per state and acoustic representation. SIVD, SDVI and SIVI decoding experiments were carried out.

Recognition rates obtained for 64 gaussians did not improve those obtained for 32 gaussians, revealing that the speech database is not large enough to use such great number of gaussians per mixtures. Moreover, as our main objective was to integrate them into a Continuous Speech Recognition system, increasing the number of gaussians could make the system too slow, so the choice of 32 gaussians per state and acoustic representation reveals as the most suitable for create a speaker recognition system.

3.3. CONTEXT DEPENDENT UNITS

Although previous results show the goodness of the database, the use of context dependent units allowed us to study the phone context distribution. Furthermore nowadays speech recognition systems are based on context dependent units since better acoustic modelization and results are achieved.

A statistical analysis of the training corpus stands to get a total of 2597 different triphones and 447 different biphones. To get enough samples to train acoustic models, only those context dependent units with more than 300 appearances were used in the experiments. In this case, the number was reduced to 278 (10,7 %) for triphones, 226 (50,6%) for right biphones and 238 (53,2%) for left biphones.

Three training procedure were carried out to train the three sets of context dependent units. Each unit in each experiment was initialized using the corresponding context independent phones already trained for previous experiments. Then the samples of each context dependent unit available in the database were used to re-estimate the HMM probability distributions. At decoding time, the phone-like models were also included to guarantee unit coverage at testing sets. Table 4 and Table 5 show phone recognition rates for discrete and continuous HMMs respectively when these sets of context dependent units were used.

As expected, better results were achieved with contextual

Units\test set	SDVI	SIVD	SIVI
Triphones	65.33	66.88	64.85
Right biphones	65.35	66.84	64.98
Left biphones	65.00	66.96	64.23

Table 4: Phone recognition rates obtained trough SDVI, SIVD and SIVI decoding experiments for three context dependent unit sets when discrete HMM were used.

	M	SDVI	SIVD	SIVI
Triphones	8	69.74	71.20	69.17
	16	71.42	72.60	70.22
	32	73.57	75.56	74.08
Right biphones	8	71.10	72.19	69.81
	16	72.44	73.66	70.85
	32	74.45	76.13	73.88
Left biphones	8	71.73	73.15	70.49
	16	73.08	74.40	71.36
	32	74.65	75.78	74.23

Table 5: Phone recognition rates obtained trough SDVI, SIVD and SIVI decoding experiments for three context dependent unit sets when continuous HMM were used with several mixture sizes (M = 8, 16 and 32).

dependent units, especially when using mixtures of 8 and 16 gaussians. However, more important improvement in system performance is expected when using context dependent models to built lexical models. One way to estimate this improvement at acoustic-phonetic decoding level is to add a phonological language model considering the compatibility between pairs of units. Preliminary results showed in Table 4 and Table 5 do not consider units compatibility.

These experiments confirm the suitability of the database design also in terms of phone context distribution. The number of different contexts as well as the number of samples of more frequent contexts is high enough to allow reductions in phone recognition errors, even in preliminary experiments

4. CONCLUSIONS

The evaluation of an acoustic database in Basque language to perform speech applications was presented. The number of samples for each unit, the phonetic balance and the phone context distribution was verified as adequate to achieve good acoustic models to be integrated in Continuous Speech Recognition systems. Thus, it may be asserted that it was correctly designed.

5. References

Esprit Project 2589 (SAM), 1991. Multi-lingual speech input/output assessment, methodology and standardization. Technical report, Esprit.

Linde, Y., A. Buzo, and R. M. Gray, 1980. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, COM-28(1).

López de Ipiña, K., I. Torres, and L. Oñederra, 1998. A speech database in basque language. In *Workshop on Language Resources for European Minority Languages*. Granada.

López de Ipiña, K., I. Torres, L. Oñederra, and L. J. Rodriguez, 2000. Selection of sublexical units for continuous speech recognition of basque. In *Proc. of International Conference of Spoken Language Processing (ICLSP), N814*. Beijing.

Mitxelena, K., 1977. *La lengua vasca*. Leopoldo Zugaza, Durango.

Rodriguez, L. J. and I. Torres, 2003. Comparative study of the baum-welch and viterbi training algorithms applied to read and spontaneous speech recognition. In *1st Iberian Conference on Pattern Recognition and Image Analysis (lbPRIA2003)*. Puerto de Andratx (Mallorca, Spain).