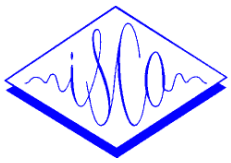

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

DISCRIMINATIVE-TRANSITIONAL/STEADY UNITS FOR SPANISH CONTINUOUS SPEECH RECOGNITION*

A. Varona[1!], I. Torres[1], F. Casacuberta[2]

[1]Dpto. Electricidad y Electrónica. Universidad del País Vasco.

Apdo. 644 - 48080 Bilbao. SPAIN. E-mail (amparo@lc. ehu.es) (manes@lc. ehu.es)

[2]Dpto. Sistemas Informáticos y Computación. Universidad Politécnica de Valencia.

Apdo. 22012 - 46071 Valencia. SPAIN. E-mail (fcn@iti.upv.es)

ABSTRACT.

The design of current acoustic-phonetic decoders for a specific language involves the selection of an adequate set of sub-lexical units. In previous works [1] a set of context independent units was presented for Spanish Continuous Speech Recognition. The aim of this work was to extend this basic set representing context variability under a discriminative-transitional/steady criterion. Our main goal was to obtain a good trade-off between discriminative ability, context variability and number of units. Finally, and within the framework of Hidden Markov Modelling, different series of acoustic-phonetic decoding experiments were carried out over a Spanish Continuous Speech corpus. The phone recognition rates obtained through these experiments showed that the proposed criterion seems to be an interesting approach to model context variability in Spanish.

1. INTRODUCTION

Most of the present ambitious Continuous Speech Recognition (CSR) systems are based on adequate acoustic modelling of chosen sets of sub-lexical units. These sub-lexical models can be used to build word models of a given vocabulary, or in an Acoustic-Phonetic Decoder (APD) to obtain the best interpretation of an input utterance in terms of a sequence of sub-lexical units. The first step in the process of designing such systems is the selection of the type of sub-lexical units. This depends on the language considered and should be made on the basis of language coverage, context variability and frequency of occurrence in the available training corpus.

The choice of phones as sub-lexical units presents several advantages. The most important is that the number of such units is low enough to obtain a high score of occurrences for each unit in the training set. Thus, the corresponding models will be well learnt. On the other hand, they are context-independent so that they could be used to model new words not appearing in the basic training set.

However, the acoustic variability of some units may not be well represented. Many times this variability is really task dependent [2]. To deal with this problem, several proposals can be found in the literature: reducing acoustic variability by increasing the number of models for each context independent unit, choosing context dependent phones [2][3], diphones [4], triphones [3], polyphones [5], context-freezing units [6], demisylables [7], etc. All these sets are based on a previously selected baseline of phone-like units.

The aim of this work was to present an adequate set of sub-lexical units for Spanish CSR selected under a discriminative-transitional/steady criterion. In Section 2, this choice is discussed and different sets of units are presented. Section 3 summarises the methodology and experimental environment used in the series of acoustic phonetic decoding experiments carried out to test the proposed set of units. These experiments and results are presented in Section 4. Finally, some concluding remarks are summarised in Section 5.

2. SELECTION OF SUB-LEXICAL UNITS.

In previous works [1] [8] a set of context-independent units was introduced for Spanish CSR. This set consisted of 23 Phone-Like Units (PLU) that roughly correspond to the 24 Spanish phonemes [9]. Table 1 shows this set using the IPA and SAM transcriptions [10] for their representation..

Table 1: Basic set of phone-like units (PLU): IPA and SAM transcriptions are used for their representation.

	IPA transcription	SAM transcription
Occlusive :	[p] [t] [k] [b] [d] [g]	p t k b d g
Nasals:	[m] [n] [h]	m n J
Fricative:	[f] [θ] [s][λ][y] [x]	f T s L,Z x
Affricate:	[tʃ]	tS
Liquids:	[l] [r] [rr]	l r rr
Vowels:	[i] [e] [a] [o] [u]	i e a o u

* Work partially supported by the Universidad del País Vasco under grant (UPV 224.3/0- EA041/93) and by the Spanish CICYT under grant (TIC 94-0210-E)

! Supported by the Basque Goberment under grant BFI94.161-AE

This small set was considered adequate for Spanish after a deep study of the language under phonetic criteria [8] and after experimenting with other proposals including larger sets of allophones [11].

However, the acoustic variability is not well represented with context independent units. As a consequence, we proposed a discriminative-transitional/steady criterion [4] to extend this basic set of PLU´s. Our main goal was to obtain a good trade-off among discriminative ability, context variability and number of units.

Table 2: First set of discriminative units (DU).

Steady units
[p] [t] [k] [b] [d] [g] [m] [ɳ] [ʧ] [β] [ð] [ɣ] [f] [θ] [s] [y] [x] [l] [r] [rr] [i] [e] [a] [o] [u]
Transitions
Diphthongs:
/ja/ /je/ /jo/ /wa/ /we/ /aj/ /ew/ /ae/ /ea/ /oa/ /oe/
Consonant groups:
/pr/ /tr/ /βr/ /dr+ðr/ /βl/
Unvoiced occlusive-vowel
/pa/ /pe/ /ta/ /te/ /ti/ /to/ /ka/ /ke/ /ki/ /ko/
Voiced consonant-vowel:
Occlusive:
/de/ /βa/ /βe/ /βi/ /βo/ /ða/ /ðe/ /ðo/ /ðu/ /ɣa/ /ɣo/ /ɣu/
Nasal:
ma/ /me/ /mi/ /mo/ /na/ /ne/ /ni/ /no/
Liquid:
/la/ /le/ /li/ /lo/ /lu/ /ra/ /re/ /ri/ /ro/ /ru/ /rre/
Vowel-voiced consonant:
/an/ /en/ /in/ /on/ /un/ /al/ /el/ /il/ /ol/ /ar/ /er/ /ir/ /or/
Voice consonant-voice consonant:
/lrr/ /nl/

Table 3: Second set of discriminative units (DU).

Steady units
[p] [t] [k] [b] [d] [g] [m] [ɳ] [ʧ] [β] [ð] [ɣ] [f] [θ] [s] [y] [x] [l] [r] [rr] [i] [e] [a] [o] [u]
Transitions
Diphthongs:
/ja/ /je/ /jo/ /wa/ /we/ /aj/ /ew/ /ae/ /ea/ /oa/ /oe/
Consonant groups:
/pr/ /tr/ /βr/ /dr+ðr/ /βl/
Unvoiced occlusive-vowel:
/pa/ /pe/ /ta/ /te/ /ti/ /to/ /ka/ /ke/ /ki/ /ko/
Voiced consonant (liquid)-vowel:
/la/ /le/ /li/ /lo/ /lu/
Others:
/rre/ /el/

The extended set of Discriminative Units (DU) includes two kinds of units: a basic collection of steady units close to the Spanish phone set and a bigger set of units representing transitions between pair of phones. The selection of the transitional units was made under a discriminative criterion. Thus, only well acoustically characterised transitions were considered: diphthongs, consonant groups, unvoiced occlusive-vowel transitions, voiced consonant-vowel transitions, vowel-voiced consonant transitions and some transitions between voiced consonants.

The frequency of occurrence of each unit in the available corpus was also considered in order to get well-trained acoustic models. Thus a set of roughly 100 units was finally selected. The IPA transcription of this set is represented in Table 2.

In order to test this proposal, a set of acoustic-phonetic decoding experiments was carried out (see section 4). These experiments showed that some of the transitions contributed to increase the confusion rate among several phones. As a consequence, the number of transitions was reduced resulting in the set showed in Table 3. As a consequence, the transitions chosen, and thus the final set of units, obviously was corpus dependent.

3. EXPERIMENTAL ENVIRONMENT

The corpus used in this work consisted of two kinds of sentences: 120 phonetically balanced sentences used in the training phase and in vocabulary dependent decoding experiments and 50 sentences obtained form current Spanish narrative used in decoding experiments. All the sentences were uttered by 10 speakers: 7 were used in the training phase and speaker dependent decoding experiments and 3 for speaker independent decoding experiments. This corpus included 48,000 PLU´s and 67,000 DU´s.

Single (SC) and Multiple Codebook (MC) Discrete Hidden Markov Models [2] were used to model each sub-lexical unit. Each unit was represented by a simple left to right topology with three states without transition between non-consecutive states and a loop at the second state.

The whole corpus was acquired at 16 KHz and parametrised, resulting in vectors of dimension 11, i.e. 10 Cepstrum Coefficients (CC) plus Energy (EN). From these parameters we have got their respective first derivatives (ΔCC and ΔEN). A codebook of 128 codewords was used for SC experiments (CC+EN). For MC experiments three codebooks (CC, ΔCC and EN+ΔEN) of 128 codewords were used.

Both sets of DU´s (Table 2 and Table 3) were evaluated and compared to the basic set of PLU´s (Table 1) in terms of phone, i.e. PLU´s, recognition rates.

For each experiment, the following values were computed: 1) the number of PLU´s that were recognised correctly (c); 2) the number of PLU´s that were inserted (i), deleted (d) and substituted (s). These values were obtained by an editing comparison between the output of

the APD and the correct PLU transcription of each test utterance.

From these values, the following parameters were obtained:

Percent Correct: $P_c = 100\ (c / i+s+d+c)$

Substitution rate: $P_s = 100\ (s / i+s+d+c)$

Deletion rate: $P_d = 100\ (d / i+s+d+c)$

Insertion rate: $P_i = 100\ (i / i+s+d+c)$

4.- EXPERIMENTS AND RESULTS

Two kinds of acoustic-phonetic decoding experiments are presented. A preliminary set of experiments was carried out to test the set of units presented in Table 2. In this case only a sub-corpus of the available corpus was used to preserve the statistical independence of the whole corpus test set. Then the finally proposed set of units (Table 3) was tested over the whole corpus test set in a second series of experiments.

4.1. PRELIMINARY SET OF EXPERIMENTS

The training sub-corpus consisted in this case in 30 sentences uttered by 4 speakers (20,000 DU's). The test sub-corpus consisted in 20 different sentences (12,000 DU's) uttered by the same speakers in Speaker Dependent (SD) experiments and by 4 different speakers in Speaker Independent (SI) experiments.

A phoneme pair-grammar with two different constraint levels was included: weak constraints (WC) and strong constraints (StC). In the WC case transitional units were not allowed to be consecutive. The StC model forced sequences of units steady/transition/steady when the transitional unit was defined.

Table 4 shows the phone recognition rates obtained through these experiments, when the basic set of PLU (Table 1) and the new set of DU units (Table 2) were used.

Table 4.-Phone recognition rates (in %) obtained with the first set of DU's (Table 2) and PLU's (Table 1)

			P_i	P_b	P_s	P_c
SD	PLU		7	10	25	58
	DU	WC	7	10	24	59
	DU	StC	5	13	29	53
SI	PLU		3	12.5	30	54.5
	DU	WC	3	12	30	55
	DU	StC	5	13.5	32.5	49

DU units slightly improved the PLU phone recognition rates only when the WC model was used. A more specific analysis of substitution errors among phones [12] would show that only a small set of voiced consonant-vowel transitions let to improve the system performance. This small set included very frequent Spanish "function words" like /la/ /el/ /le/ /lo/. Diphthongs are also very frequent in Spanish and their formant transitions were well characterised by these transitional units.

Unvoiced occlusive were better identified when their transitions with the next vowel were included. Finally the specific Spanish sounds /r/ and /rr/ (simple and multiple vibrant liquids) are very difficult to characterise by themselves due to the absence of any specific spectral feature [9]. Thus consonant groups and some vowel transitions including those phonemes needed to be considered. As a consequence, a more reduced set of DU units was then proposed (see Table 3).

4.2 FINAL EXPERIMENTS

The object of these experiments was to test the final set of DU's (Table 3) in more exhaustive decoding experiments.

The whole training corpus consisted of 840 utterances that correspond to 120 phonetically balanced sentences uttered by 7 speakers (4 female and 3 male), resulting in a total of 48,000 PLU's and 67,000 DU's.

Table 5.-Summary of the phone recognition results (in %)

			P_i	P_b	P_s	P_c
SD VI	SC	PLU	3	14	25	58
		DU	3	15	24	58
	MC	PLU	3	11	22	64
		DU	3	10	21	66
SI VD	SC	PLU	3	15	27	55
		DU	3	15	27	55
	MC	PLU	4	11.5	23	61.5
		DU	4	11	22	63
SI VI	SC	PLU	3	15	28	54
		DU	3	16	27	54
	MC	PLU	3	12	24	61
		DU	3	12	22.5	62.5

Three series of experiments were carried out: Speaker-dependent vocabulary-independent (SDVI), speaker-independent vocabulary-dependent (SIVD) and speaker-independent vocabulary-independent (SIVI).

SDVI: 350 utterances that corresponded to 50 sentences from narratives uttered by 7 speakers (the same speakers of the training set), for a total of 13,000 PLU's and 16,000 DU's.

SIVD: 360 utterances that corresponded to 120 phonetically balanced sentences uttered by 3 speakers not included in the training corpus (2 male and 1 female), for a total of 9,300 PLU's and 11,500 DU's.

SIVI: 150 utterances that corresponded to 50 sentences from narratives uttered by 3 speakers not included in the training corpus (2 male and 1 female), for a total of 5,600 PLU´s and 7,000 DU´s

Table 5 shows the phone recognition rates obtained for SC and MC experiments when the PLU´s (Table 1) and the second DU´s (Table 3) sets of units were used. In this case only WC phonological model was considered for DU´s.

A slight improvement of system performances can only be observed when dynamic features were considered (MC).

5. CONCLUDING REMARKS.

The use of DU units improved the phone recognition rates in all the experiments carried out. However, the error reduction was more significant when multiple codebook was considered, up to 4%. In fact, most of DU's represent speech signal transitions and thus, correspond to very small segment of signal. As a consequence, the number of such units needed to train the models is very high, and then a larger Spanish database is required. In spite of this problem, the discriminative-transitional/steady set units seems to be an interesting approach to model context-dependent acoustic variability in Spanish, when any specific task provided

REFERENCES

[1] I.Torres, F. Casacuberta. "Spanish Phone Recognition using Semicontinuous Hidden Markov Models". *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing* (1993), vol. II, pp. 515-518.

[2] C.H. Lee, L.R. Rabiner, R. Pieraccini and J.G. Wilpon. "Acoustic modeling for large vocabulary Speech Recognition". *Computer Speech and Language* (1990) 4, pp. 127-165.

[3] K.F. Lee. "Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech recognition". *IEEE Trans. Acoust., Speech, Signal Processing* (1990) vol. ASSP -38, pp. 599-609.

[4] M. Cravero, R. Pieraccini, F. Raineri. "Definition and Evaluation of Phonetic Units for Speech Recognition by Hidden Markov Models". *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (1986), pp. 2235-2238.

[5] H. Niemann,. E. Nöth,. E.G. Schukat-Talamazini, A. Kiessling, R. Kompe, T. Kuhn, K. Ott,. S. Rieck. "Statiscal Modeling of Segmental and Suprasegmental Information ". En NATO-ASI. " News Advanced and Trends in Speech Recognition and Coding ". Granada (Spain). pp 237-260. 1993

[6] E.G. Schukat-Talamazini,. H. Niemam, W. Eckert., T. Kuhn., S. Rieck. "Acoustic Modeling of Subword units in the ISADORA speech Recognicer ". *Procc. IEEE. Int. Conf. Acoust . Speech and Signal Processing* (1993).Vol I, pp 557-580.

[7] E. Lleida, J.B. Mariño, C. Nadeu, J. Salavedra. "Demisyllable-based HMM spotting for continuous speech recognition". *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (1991), pp. 709-712.

[8] J. Llisterri. "Criterios para la elaboración de una base de datos para el reconocimiento del habla en español". ALBAYZIN Technical Report, 1991.

[9] A. Quillis. "Fonética Acústica de la Lengua Española". Ed. Gredos. 1988.

[10] SAM. "Multi-lingual Speech input/output assessment, methodology and standardisation". Esprit project 2589 (SAM). ESPRIT Technical repport, 1991.

[11] I. Torres, F. Casacuberta, L. Sánchez. "Linguistic Decoding of Continuous Speech with Hidden Markov Models".*Advances in Patteen Recognition ansd Applications* pp 207-217, Valencia, 1992.

[12] A. Varona. "Segmentacion y Seleccion de unidades discriminantes transitorias y estacionarias en Decodificación Acústico-Fonética de discurso continuo". Tesina de licenciatura. Noviembre 1993.