THE DEMIPHONE: AN EFFICIENT SUBWORD UNIT FOR CONTINUOUS SPEECH RECOGNITION [1]

Jos B. Mari o, Albino Nogueiras, Antonio Bonafonte
Universitat Polit cnica de Catalunya
c) Jordi Girona 1-3
08034 Barcelona
SPAIN
{canton / albino / antonio}@gps.tsc.upc.es

ABSTRACT

In this paper we introduce the demiphone as a contextual phonetic unit for continuous speech recognition. A phone is divided into two parts: a left demiphone that accounts for the left side coarticulation and a right demiphone that copes with the right side context. This new unit discards the dependence between the effects of both side contexts, but provides a better training of the transition between phones. The demiphone can be seen as a heuristic clustering of states that allows a more smoothed training of hidden Markov models and additionally supplies a simple way to create unseen triphones. We report experimental evidence that demiphones outperform the usual combination of triphones, right-side and left-side biphones and monophones.

1. INTRODUCTION

Acoustic modeling for continuous speech recognition is a topic under permanent research, because the performance of a speech recognition system greatly depends on the acoustic modeling quality. Hidden Markov models (HMM) of phones are the most popular option for modeling speech sounds. With these models and by means of a phonetic transcription it is easy to modelize the words in the vocabulary of the task to be recognized. In order to cope with the coarticulation effects on the realization of phonemes, context dependent phonetic units have been defined. Thus, triphones (TRPH) have been proposed to take into account both contexts of a phoneme, the previous and posterior phonemes. A simpler unit is the biphone, a phone that depends only on one context, the left side phoneme (LBPH) or the right side one (RBPH). The recognition systems that incorporate these types of subword units clearly give better performance than systems designed with context independent phones (CIPH) only. Their main drawback stems from the huge amount of speech material necessary to train context dependent units; or, in other words, the difficulties arise from the lack of material to train some HMMs.

Particularly, task independent modeling has received the attention of researchers. The goal is to obtain acoustic models of speech from a general (phonetically balanced) database and use them in a task oriented recognition system. This approach tries to save the cost of a task dependent speech database without significant loss of performance. The main problem to be solved is the mismatch between the set of phonetic units that can be trained from the phonetically balanced data base and the set of units necessary to modelize the target vocabulary.

In order to overcome the limited size of data bases, some relatively successful techniques have been proposed. Clustering of models or states reduces the number of parameters to be learnt and provides more robust (smoothed) estimates. The design of decision trees to steer the clustering procedure can yield a straightforward way to provide a model for an unseen phonetic unit [1]. Alternatively, it has been proposed to build triphones from parts of biphones which are more easily trainable [2]. The first states of the triphone model are taken from the beginning of a left-side biphone model, and the final states are borrowed from a right-side biphone. The main drawback with this method is that both pieces are trained independently without paying any attention to the future union. For instance, the most phonetically suitable point where the juncture could be done is not learnt from data.

In this paper a new phonetic unit is introduced: the demiphone. This unit shares in a simple way the advantages of clustering (or tying) of states with the ability of generating unseen context dependent units. The paper is organized as follows. An overview of the speech data bases and the recognition system is presented in the next section. The demiphone is defined and theoretically supported in Section 3. Afterwards, Section 4 reports the results of the experiments carried out to evaluate the performance of the introduced demiphone. A discussion follows in Section 5. The paper ends by remarking the most important conclusions and advancing future work.

2. EXPERIMENTAL FRAMEWORK

2.1 Databases

In order to test the new unit, we have accomplished a task and speaker independent training with part of the EUROM.1 [3] Spanish material (43 speakers with a total of 842 utterances) and other additional speech recorded in our laboratory. The overall training database gathers 1529 utterances from 57 speakers.

Three different tests were recognized:

[1] This research was supported by the CICYT under contract TIC95-0884-C04-02

a) In order to assess the phonetic recognition (PHR) performance of the demiphone we use 700 utterances from 33 speakers.

b) The next tests are taken from two application tasks. The first one is constituted by oral inquiries into a geographic information database (GDQ) [4]. The vocabulary has 310 words, the average number of words for sentence is greater than 9 and the test set perplexity of the task assessed by bigrams is 12. We use 464 utterances from 12 speakers as test material.

c) The second application test is formed by 1161 orders to TELEMACO from 53 users. TELEMACO [5] is a system for automatic voice dialling based on the recognition of commands in fluent speech. The set of command words is composed of the digits and fifteen dialling words (call, answer, transfer, number, etc.). An order can require from one to nine command words and includes extraneous words.

The training material and the PHR and GDQ tests were recorded originally with a sampling frequency of 16 kHz. The signals from TELEMACO were recorded a 8 kHz; consequently, the training material was downsampled to obtain an 8 kHz version.

2.2 System overview

The speech was parameterized with mel-cepstrum coefficients. CMS (cepstral mean substraction) was used since the training and testing data bases were recorded differently. First and second order differential parameters plus the differential energy were employed.

The recognition system models the phonetic units by gaussian SCHMM with quantization to the 6 (2 for the energy) closest codewords. The size of the codebooks was 256 (64 for the differential energy) when processing 16 kHz speech and 128 (32) was used with 8 kHz material.

In the PHR experiment the system incorporates a grammar to allow only the concatenation of units for which contexts agree. The GDQ task is modeled by an X-gram [6] that yields a test perplexity of 8 with 648 states.

The system decodes each utterance of TELEMACO in commands and fillers. The command models were built with the phonetic units analyzed in this paper. The fillers are models of clusters of syllables with 8 states and no skips. Both types of models were estimated from the 8 kHz version of the training speech.

3. THE DEMIPHONE

3.1 The definition

Results from recent works [2,7-10] seem to support the following provisional conclusions:

a) Only in a very few cases coarticulation variants depend on both the left and the right contexts. Triphones give a very reduced improvement in performance, if any, over that reached with left and right side biphones [7-8].

b) In most of the cases coarticulation effect on one side of the phone is practically independent of the other. Triphones built with parts of biphones (the first state from a left-side biphone and the rest from a right-side biphone) exhibit an excellent behaviour [2]. Tying the left states of the triphones that share the same left context (and equivalently for the right states) provides satisfactory acoustic modeling [9-10] in speaker dependent systems.

As a consequence, we propose a new subword unit: the demiphone (DPH). A phone is conceptually divided into two parts: a left part that corresponds to the beginning of the phone and encompasses the left side coarticulation variations, and a right part that does the same mission for the final part of the phone. Thus, we distinguish two types of demiphones: left side demiphones (LDPH) and right side demiphones (RDPH). As an example, the Spanish word "osa" is transcribed with demiphones in the following way: F-o, o+s, o-s, s+a, s-a, a+F. The units F-o, o-s and s-a are left side demiphones; o+s, s+a and a+F are right side demiphones. The symbol F denotes the boundary of a word; we do not consider interword contexts yet. The introduction of the demiphone unit has useful advantages:

a) Phones of the task for which left and right side contexts are unseen together in training can be modeled in a natural way during recognition. A simple phonetic transcription solves the situation. It is not necessary to build a new (triphone) model artificially.

b) Both left and right side coarticulation variations are modeled. Thus, the modelization of the most relevant context is guaranteed without using triphones or paralleling left and right side biphones.

c) The number of demiphones saturates much faster than the number of triphones. In Table I we show the number of triphones and demiphones that appear in the training corpus a number of times over a given threshold. As a consequence, the percentage of speech material available to train hidden Markov models (coverage) is much higher for demiphones than for triphones.

d) The training material is efficiently used for modeling left and right contexts.

e) The training and recognition algorithms are simplified in comparison with the tying alternative.

	triphone		demiphone	
threshold	number	coverage	number	coverage
200	46	29%	159	75%
100	101	43%	254	88%
50	242	59%	384	95%
25	538	77%	476	98%
10	1078	91%	574	99%
1	2137	100%	690	100%

Table I.- Number of triphones and demiphones and coverage in the training material as a function of the counting threshold.

f) If we choose to model a demiphone with half the number of states dedicated to a phone, the number of parameters is reduced.

3.2 Sets of units evaluated

In this paper we compare the recognition performance attained by the demiphone and a classical set (TRL) of context dependent phone units formed by triphones, right-side and left-side biphones and context independent phones [11]. We distinguish 25 phonemes for Spanish and we trained only the demiphones with at least a given number N of realizations in the test; the rest of demiphones were merged in a unique left demiphone and a unique right demiphone for every phone. Thus, we have 50 context independent demiphones (CIDPH). The TRL set was defined in the following way: we modeled the triphones with N appearances or more in the training corpus, with the rest of the material we trained the RBPH units that surpass the threshold N; afterwards, on the remaining data the LBPH units were estimated and, finally, the 25 CIPH were added to get a 100% coverage.

In order to choose a suitable threshold N, we carried out a recognition experiment. We tried three values for N (50, 100 and 200). After training the corresponding units, their performance with the geographic data query task was evaluated. In view of the results (in Table II), we selected N=100. The forthcoming results will always refer to this value.

Table III provides the composition of the TRL and DPH sets. Additionally, the number of hidden Markov model states to be trained is indicated. Every phone in the TRL set is built with four states. The demiphone set has two states for each unit; however, the structure of the model is different for the left and the right demiphones: the model of the left demiphone can be abandoned from the first state; on the contrary, the two states of a right demiphone must be visited. In this way we reproduce as closely as possible the structure we use for phone models (a four state model where one skip is allowed during transitions between states).

Table III also shows the same information about the so-called DPtrl set, which is considered for discussion purposes. This set is composed of the demiphones necessary to build the TRL set. It is important to remark that the DPtrl set emulates the TRL units only. No generation of new units is allowed. As can be seen in Table III, the main difference between the TRL (or Dptrl)

Unit	50	100	200	Gram
TRL	41.3	45.8	45.5	NO
DPH	50.8	53.2	50.9	NO
TRL	92.3	94.7	93.5	X
DPH	95.1	95.3	95.0	X

Table II.- Word accuracy, as defined in (1), for the GDQ task with NO grammar and X-gram as a function of the unit counting threshold.

TRL set	
TRPH	101
RBPH	111
LBPH	14
CIPH	25

DPH set	
RDPH	123
LDPH	131
CIDPH	50

DPTRL set	
RDPH	123
LDPH	64
CIDPH	50

number of states	
TRL	1004
DPH	608
DPTRL	474

Table III.- Contribution of the different context dependent units to the evaluated phonetic sets and overall number of states for every set.

test	total	in TRL	coverg.	gen. by DPH	coverg.
PHR	2013	101	41%	775	77%
GDQ	573	98	40%	358	77%

Table IV.- Coverage of the test by triphones when either the TRL set or the DPH units are used. The number of different triphones existing in the test, the ones provided by the TRL set and the ones generated by demiphones are all included.

and the DPH sets is the capability to cope with the left-side coarticulation. Clearly, the demiphone collection is the best prepared to deal with it.

Finally, Table IV illustrates the simultaneous coverage of both left and right side coarticulation provided by triphones and demiphones for the PHR and GDQ tests. Whereas the TRL set has only one hundred triphones to offer, the DPH set can generate several hundreds of triphones. As for the TELEMACO command vocabulary, it is worth mentioning that only 35% of phones are modeled by triphones of TRL, whereas 75% is covered by triphones generated by demiphones.

4. RESULTS

Table V and Table VI show the performance reached in the test experiments where the phonetic units are less helped by language modeling: phonetic recognition (PHR) and word recognition without task grammar. The figures reported are the following:

C = percentage of correct recognitions
S = percentage of substitutions
D = percentage of deletions
I = percentage of insertions
A = accuracy = C/(1+I/100) (1)

Table VII shows the word accuracy and the percentage of sentences correctly recognized in both GDQ and

Unit	C	S	D	I	A
TRL	78.2	16.2	5.6	9.5	71.4
DPH	78.5	15.2	6.3	7.0	73.4
DPtrl	77.7	15.8	6.5	7.5	72.3

Table V.- Phonetic recognition performance.

Unit	C	S	D	I	A
TRL	62.9	29.2	7.9	37.4	45.8
DPH	65.2	27.6	7.2	22.6	53.2
DPtrl	61.8	30.6	7.6	40.1	44.1

Table VI.- Word recognition scoring for the Geographic Data Query task without grammar.

	GDQ task		TELEMACO	
Unit	A	S	A	S
TRL	94.7	72.4	84.0	78.0
DPH	95.3	75.9	93.1	90.0
DPtrl	94.7	72.0	-	-

Table VII.- Word accuracy (A) and percentage of correctly recognized sentences (S) in the application tasks.

Unit	A	S
TRL	96.5	80.6
DPH	97.3	84.7

Table VIII.- Word accuracy (A) and percentage of correctly recognized sentences (S) in the GDQ task when using X-gram and classes of words.

TELEMACO application tasks. Finally, Table VIII reports the same figures when the X-gram for GDQ task includes classes of words decreasing the test perplexity to 6 (1206 states are needed).

5. DISCUSSION

From Tables V, VI and VII we can see that the DPtrl set performs slightly worse than the TRL units. In fact, the independent training of left and right side context seems to produce some degradation of recognition power. However, this reduction is very small. Furthermore, it is more than compensated by the smoothing and generation capability of the demiphone, as we can ascertain from the results yielded by the DPH set.

Demiphones accomplish a balanced modeling of both left-side and right-side coarticulation, since the number of units dedicated to one or other context is almost the same. Consequently, demiphones provide a better modeling of transitions between sounds. For instance, the transition inside the diphtong /jo/ is described by the concatenation j+o j-o. On the contrary, when the TRL set is utilized, this transition is well accounted for only by the links between the following units

j+o j-o+n
T-j+o j-o+F

In the rest of contexts where the diphtong /jo/ could appear, the vowel /o/ can only be modeled by a RBPH (o+any). So, there is no transition modeling.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have introduced the demiphone to model the coarticulation produced by the neighboring sounds. We hope to have provided evidence that the demiphone deserves to be considered an alternative to context dependent phones. It offers: a) the capability of coping with both left and right side contexts not simultaneously seen during training; and b) an important reduction in the number of parameters to estimate.

Nevertheless, the demiphone must be compared with triphones smoothed and generalized by decision trees, because a priori they are the most powerful tool for coarticulation description. Furthermore, the demiphone itself can be smoothed and generalized. Both tasks are our following interest.

7. ACKNOWLEDGMENTS

The authors want to thank Gustavo H. Abrego for carrying out the experimentation with TELEMACO. We also thank Eva Guijarro, Alex Risso and David Conejero for providing the grammars used in the GDQ and PHR testing.

8. REFERENCES

[1] M-Y Hwang et al, "Predicting Unseen Triphones with Senones", *IEEE Trans. Speech Audio Processing*, vol.4 n° 6, pp. 412-419, 1996.

[2] C-H Lee et al, "*A study on task-independent subword selection and modeling for speech recognition*", Proc. ICSLP96, pp. 1820-1823, Philadelphia, 1996.

[3] A. Moreno, "EUROM.1 Spanish Database". *Esprit Technology Assesment in Multilingual Applications*, Esprit Project 6919, Report D6.

[4] F. Casacuberta et al, "*Development of Spanish Corpora for Speech Recognition Research*", Proc. of Workshop on International Cooperation and Standarization of Speech Databases and Speech I/O Assesment Methods, Chiavari, 1991.

[5] E. Lleida, J.B. Mari o and A. Moreno, "*TELEMACO: a Real Time Keyword Spotting Application for Voice Dialling*", Proc. EUROSPEECH93, pp. 1801-1804, Berlin, 1993.

[6] A. Bonafonte and J.B. Mari o, "*Language Modeling using X-grams*", Proc. ICSLP96, pp. 394-397, Philadelphia, 1996.

[7] L. Fissore et al, "*Incremental Training of Speech Recognition for Voice Dialling-by-Name*", Proc. ICSLP94, pp. 447-450, Yokohama, 1994.

[8] L. Villarrubia et al, "*Context-dependent units for vocabulary-independent Spanish speech recognition*", Proc. ICASSP96, Atlanta, 1996.

[9] L. C. Wood et al, "*Improved Vocabulary-Independent Sub-Word Modelling*", Proc. ICASSP91, pp. 181-184, Toronto, 1991.

[10] J. J-X Wu et al, "*Modeling context-dependent phonetic units in a continuous speech recognition system for Mandarin Chinese*", Proc. ICSLP96, pp. 2281-2284, Philadelphia, 1996.

[11] A. Bonafonte et al, "*Study of subwords units for Spanish speech recognition*", Proc. EUROSPEECH95, pp. 1607-1610 , Madrid, 1995.