THE DEMIPHONE VERSUS THE TRIPHONE IN A DECISION-TREE STATE-TYING FRAMEWORK

José B. Mariño, Pau Pachès-Leal and Albino Nogueiras

Universitat Politècnica de Catalunya
Jordi Girona 1-3, 08034 Barcelona, SPAIN
(canton/paupag/albino)@gps.tsc.upc.es

ABSTRACT

The performances of the demiphone (a context dependent subword unit that models independently the left and the right parts of a phoneme) and the triphone are compared. Continuous density hidden Markov modeling for both types of units is tested with the HTK software using decision-tree state clustering. The speech material is taken from the SpeechDat Spanish database, composed by continuous speech utterances recorded through the public telephone network. The training corpus is speaker and task independent. Two testing sets are tried: isolated words corresponding to speaker names, city names and phonetically rich words; and numbers of Spanish identification cards and dates. The main conclusion is that the demiphone simplifies the recognition system and yields a better performance than the triphone. This result may be explained by the ability of the demiphone to provide an excellent tradeoff between a detailed coarticulation modeling and a proper parameter estimation.

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 have been proposed to take into account both contexts of a phoneme, the previous and posterior phonemes.

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 database and the set of units necessary to model the target vocabulary. In order to overcome the limited size of databases, 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. Furthermore, the design of decision trees to steer the clustering procedure can yield a straightforward way to provide a model for an unseen phonetic unit.

Recently [1], the authors have introduced the demiphone as a new contextual subword unit. A demiphone is either a left or a right part of a phone. This unit shares in a simple way the advantages of clustering (or tying) of states with the ability of modeling unseen together left and right contexts. We have reported experimental evidence that demiphones outperform the usual combination of triphones, right-side and left-side biphones and monophones. In the present work, the demiphone is compared with triphones smoothed and generalized by decision-tree state-tying, accepted as the most powerful tool for coarticulation modeling at the present state of the art.

The paper is organized as follows. Section 2 includes the description of the speech database used in our experimentation and an overview of the training and testing procedures developed with the recognition system. The demiphone is reviewed in Section 3. Afterwards, Section 4 reports the results of the comparative study between the triphone and the demiphone. A discussion follows in Section 5. The paper ends by summing up the most important conclusions and advancing future work.

2. EXPERIMENTAL FRAMEWORK

2.1 The recognition system

The recognition system (HTK [2]) used is based on continuous density hidden Markov models of phones. Monophone models were initialized via flat start training and after several embedded training reestimations, triphones were cloned from monophones, reestimated, state-level tied with decision-tree based clustering algorithms and again reestimated. Demiphones, on their part, were cloned from halves of phones, and went through the same customary cycle of reestimation, state-level tying and reestimation. During the learning of the decision-trees, the splitting of clusters was controlled by a threshold on the improvement in homogeneity. So, state-tying with different number of clusters was obtained. A standard parameterisation of 12 MFCC’s coefficients along with their first and second order differences was used, dynamic energy coefficients were also part of the acoustic vector. As a simple processing for telephone speech, cepstral mean substraction over the whole utterance was applied. The decoder used was the standard HTK decoder.

This research was supported by the CICYT under contract TIC95-0884-C04-02.

	words	number of utterances	speakers
training sentences	3993	4553	680
isolated words	1452	1992	590
names	228	316	316
cities	245	573	429
words	979	1103	516
strings of numbers	72	201	201
dates	67	402	313

Table 1. Number of different words, utterances and speakers for every corpus.

2.2 The speech databases

The speech material used in our experimentation comes from the Spanish corpus of the SpeechDat project [3]. The utterances were recorded through the public telephone fixed network, sampled at 8 kHz and quantified by the A-law at 8 bits per sample. Table 1 summarizes the most relevant characteristics of the training and testing sets.

As training material we have used phonetically balanced sentences uttered by 680 speakers from four dialectal zones and including more than 236000 phones.

We have designed three different and partly spontaneous speech tests. The first one is composed of isolated utterances with names of speakers ("names"), names of cities ("cities") and phonetically rich words ("words"). Utterances of strings of numbers corresponding to Spanish identification cards form the second test; the sentences have more than ten words in average. The third set is composed by dates and includes function and very confusing words. No grammar is provided to the recognition system when processing string of numbers or dates.

3. THE DEMIPHONE

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 and right side demiphones. 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 of /o/, /s/ and /a/, respectively. 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. A triphone can be obtained concatenating two demiphones; for instance, the triphone o-s+a is built by o-s and s+a.

As a consequence of the definition, the number of demiphones saturates much faster than the number of triphones. In Table 2 we show the number of triphones and demiphones that appear in the speech corpora. Studying the training material we found 2963 different triphones and only 841 demiphones. Consequently, more robust training can be provided to the demiphone models. Moreover, it is important to point out that the triphones in the training corpus add up less than a half of the overall set of possible triphones in Spanish. On the other hand, the seen demiphones exceed the 88% of the Spanish demiphone set. Table 2 also shows the number of units unseen during the training for every test. The contextual coverage of the tasks provided by the demiphone is clearly higher.

	triphones		demiphones	
	total	unseen	total	unseen
training sentences	2963	-	841	-
isolated words	2951	767	885	68
strings of numbers	185	15	199	0
dates	234	11	248	0

Table 2. Total number of different triphones and demiphones for every corpus. The number of unseen units in the training set for every test is also included.

Apparently, the main drawback of the demiphone is the underlying assumption that the coarticulation effect on one side of the phone is practically independent of the other or, at least, it may be modeled separately. However, results from recent works seem to support that only in very few cases coarticulation variants depend on both the left and the right contexts:

a) Triphones give a rather reduced improvement in performance, if any, over that reached with left or right side biphones [4, 5].
b) 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 behavior [6].
c) Tying the left states of the triphones that share the same left context (and equivalently for the right states) provides satisfactory acoustic modeling [7, 8] in speaker dependent systems.

On the other hand, the demiphone preserves the modelization of the transition between sounds. For instance, the demiphone o+s is always followed by o-s. Therefore, the junction of the two units models the transition between /o/ and /s/ (similarly as the diphone used to synthesize speech does). This is an interesting property of the demiphone, because it is known that transitions convey a great deal of speech inteligibility.

Theoretically, the triphone offers a better modeling of contexts, but in practice at the expense of a very much higher number of parameters and a lower contextual coverage of the application tasks. Decision-tree state-tying [9] is a powerful strategy that attempts to overcome these disadvantages by sharing the states of the hidden Markov models that belong to the triphones corresponding to the same sound. Thus, the number of parameters to be estimated is reduced and a more robust training is achieved. Additionally, the decision-tree learnt during training provides a way to generate models for unseen units. At this point, we are faced with two approaches that represent different tradeoffs between, on the one hand, smoothing and generalization capability, and on the other hand, power of coarticulation modeling. The next section is dedicated to gaining insight into this question.

4. EXPERIMENTAL RESULTS

In this experimental work 31 allophones were distinguished for the Spanish language. The approximant allophones /B/, /D/ and /G/ of the voiced plosive consonants were added to the basic set of 25 sounds. The voiced palatal fricative /jj/ and the voiced palatal lateral /L/ were considered separately. The velar nasal consonant /N/ was also modeled. Furthermore, the voiced allophone /z/ of /s/ was included.

In the summary of results that follows the recognition scoring is provided in percentage of utterances recognized correctly for the isolated words (the average of the three sets) and in accuracy of recognized words for strings of numbers and dates.

4.1 Preliminary experiments

In order to obtain a first reference, we have experimented with 3-state hidden Markov models for the triphone and 1 gaussian for state. Decision-tree clustering was conducted in two versions. The first one was state dependent as suggested in [10], i.e., the first states of the models are tied separately of the second and third states, central states do not share tying with extreme states and so on. This strategy is reported because it has conceptual similarities with the definition of the demiphone. The clustering was also carried out without state constraints. Under the label "state dependent" Table 3 exhibits the performance for the first strategy. The column "state independent" corresponds to the unconstrained clustering. The total number of states for each type of training is also included. As in the rest of the paper, these figures refer to the tying that provides the optimum scoring among the several tying options estimated during the training. As may be noticed, when an important acoustical mismatch exists under a task independent phonetic training, the state dependent strategy does not work as well as it does for task dependent designs.

	state dependent	**state independent**
isolated words	86.0	88.3
strings of numbers	75.6	75.2
dates	63.8	63.8
states	1276	2413

Table 3. Optimum triphone recognition results with 3-state models and 1 gaussian per state. Decision-tree tying shares all the states of the models for a same phone or is state dependent. The overall number of states is also provided.

	triphones	**demiphones**
isolated words	89.6	90.3
strings of numbers	77.1	77.4
dates	65.5	66.7
states	4170	1273

Table 4. Optimum recognition performance with decision-tree tying, 4 states per phone and 1 gaussian per state. The overall number of states is also included

From now onwards, the number of states is fixed to four for the triphones (where one skip is allowed during transitions between states) and two for the demiphones. 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 whereas the two states of a right demiphone must be visited. In this way we reproduce as closely as possible the structure used for triphone models. The training was again implemented with decision-tree state-tying. No state dependence was forced. For the best tying, Table 4 shows the performance achieved with triphones and demiphones and 1 gaussian per state. The demiphone provides a better performance while requiring less than a third part of the parameters.

4.2 Modeling with a mixture of 3 gaussians per state

Once the decision trees were learnt modeling the continuous density in every state with 1 gaussian, a new training was accomplished with a mixture of 3 gaussians per state. The best recognition performance achieved is shown in Table 5. For the triphone set, the performance corresponding to a design with a number of parameters close to the number of parameters required for the demiphones is provided. As main conclusions we may mention:

a) For every testing corpus, the performance of the demiphone improves more than that of the triphone.

b) In the case of the triphone, a greater complexity to model the continuous densities of the states requires a reduction in the number of states. This fact suggests a near saturation behavior. In fact, six gaussians per state yield a no noticeable improvement in performance.

c) The advantage of the demiphone over the triphone is incremented when designs with the same number of parameters are considered. As an exception, when the "string of numbers" corpus is tested, a reduction in the number of states is followed by an improvement in performance. Neither this task nor the "dates" task is very demanding as regards the number of different units so the pooled estimation provided by tying seems to be beneficial. However this does not hold for the "dates" task. The authors cannot provide a satisfactory explanation for it.

	triphones		**demiphones**
	optimum	**equivalent**	**optimum**
isolated words	91.6	90.7	92.8
strings of numbers	77.6	79.5	80.6
dates	69.7	69.1	71.0
states	1904	1246	1273

Table 5. Optimum recognition performance with 3 gaussians per state. For the triphone, the set with an equivalent number of parameters is included. The overall number of states is also supplied.

4.3 The benefits of state-tying

In our initial study of the demiphone [1] we selected the trainable units via a previously established threshold N. We trained only the demiphones with at least N realizations in the training corpus. The rest of demiphones were merged in a unique left demiphone and a unique right demiphone for every phone. The triphone set was determined similarly. The triphones with N appearances or more in the training corpus were modeled; the rest of the material was dedicated to train right biphones that surpassed the threshold N; afterwards, on the remaining data left biphones were estimated and, finally, monophones were added to get a 100% coverage. Now, we reproduce this selection of units with N=100 and the present training material. Under the columns “≥100”, Table 6 shows the performance achieved with triphones and demiphones and 3 gaussian per state. The recognition scoring for the best tying configuration is reproduced from table 5. Some interesting observations can be made:

a) As could be expected, the state-tying improves the performance of triphones significantly.
b) This is not the case for the demiphone. Table 7 shows the percentage of material that can be modeled (coverage) by the contextual units defined by threshold. It is clear that the smoothing and generalization capabilities of the tying algorithm have little to offer to the demiphone to cope with the strings of numbers and dates corpora.
c) Demiphones selected by threshold compare favourably with triphones estimated by state-tying.

	triphones		**demiphones**	
	≥ 100	**tying**	**≥ 100**	**tying**
isolated words	86.6	91.6	91.3	92.8
strings of numbers	76.5	77.6	80.0	80.6
dates	66.3	69.7	70.8	71.0
states	2808	1904	1182	1273

Table 6. Optimum recognition performance with 3 gaussian per state. The set of units was established by threshold (≥ 100) or decision-tree tying was implemented. The overall number of states is also included.

	triphones	**demiphones**
training sentences	72.7	97.6
isolated words	43.5	91.4
strings of numbers	69.6	96.7
dates	67.4	98.5

Table 7. Coverage of the training and testing corpora by the trainable contextual triphones and demiphones selected by threshold.

5. DISCUSSION

The reported results support that independent modeling of left and right contexts is a successful strategy to cope with coarticulation. It yields an improvement when the phonetic mismatch between training material and task is important and when it is small. Thus, the introduction of the demiphone is justified. It could be argued that the demiphone is equivalent to a priori and heuristic state-tying. Truly, it can be seen that way. However, the demiphone offers a very simple interpretation and an easy phonetic transcription of speech. In our opinion, the demiphone deserves for these reasons to be considered a phonetic unit itself.

Furthermore, the demiphone needs little tying, if any. For instance, the initial 1682 states of the demiphones are reduced after tying to 1273, i.e., in average 4 initial states are represented by 3 final states. On the contrary, the initial 11852 states of the triphone are merged in 1904 states, i.e., 6 initial states are substituted by only 1 final state. Consequently, the demiphone represents a determined context more precisely than the triphone does and the modeling of transitions between sounds is better accomplished.

It is worth mentioning that the coverage capability of the demiphone is particularly suited to be profitable in a discriminative training approach, where unseen units cannot be directly modeled.

6. SUMMARY

The demiphone has been compared with the triphone in a decision-tree state-tying framework. The demiphone and the triphone represent two different tradeoffs between, on the one hand, smoothing and generalization capability, and on the other, power of coarticulation modeling. The reported experimental evidence shows the advantages that demiphones supply:

- The recognition performance of demiphones is better than that of triphones.
- Demiphones require less parameters than triphones.
- As demiphones selected by threshold provide a satisfactory performance, state-tying may be no necessary.

Further research will be done on the study of the demiphone in a discriminative training framework.

7. REFERENCES

1. José B. Mariño et al., “The demiphone: an efficient subword unit for continuous speech recognition”, *Proc. EUROSPEECH97,* pp.1215-1218.
2. S. Young et al., *The HTK book (for version 2.0),* March 1996.
3. A. Moreno, R. Winsky, “Spanish Fixed Network Speech Corpus”, *SpeechDat Project LRE-63314*.
4. L. Fissore et al., “Incremental Training of Speech Recognition for Voice Dialling-by-Name”, *Proc. ICSLP94,* pp. 447-450.
5. L. Villarrubia et al., “Context-dependent units for vocabulary-independent Spanish speech recognition”, *Proc. ICASSP96,* pp. 451-454.
6. C-H Lee et al., “A study on task-independent subword selection and modeling for speech recognition”, *Proc. ICSLP96,* pp. 1820-1823.
7. L. C. Wood et al., “Improved Vocabulary-Independent Sub-Word Modeling”, *Proc. ICASSP91,* pp. 181-184.
8. 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.
9. M-Y Hwang et al, “Predicting Unseen Triphones with Senones”, *IEEE Trans. on Speech Audio Processing, Vol. 4 nº 6, 1996, pp. 412-419.*
10. C. Chesta et al., “Bottom-up and top-down state clustering for robust acoustic modeling”. *Proc. EUROSPEECH97,* pp.11-14.