Validating Different Flexible Vocabulary Approaches on the Swiss French PolyPhone and PolyVar databases

Andrei Constantinescu, Olivier Bornet, Gilles Caloz and Gérard Chollet

Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)
C.P. 592, CH-1920 Martigny, Switzerland

ABSTRACT

In this paper, we attempt to validate the flexible vocabulary approach for speaker independent isolated word and connected words recognition. We compare the performance of classical whole word HMMs against different sets of subword units. For this purpose, we model phonemes, diphones and words of the (Swiss) French language. The recognition rates obtained with phoneme models are monitored as we increase the amount of training data. The results of the described experiments validate the flexible vocabulary approach and show advantages and disadvantages of both propposed subword units against common whole word HMMs.

1. INTRODUCTION

Subword units represented as Hidden Markov Models (HMMs) [6], is to be one of the most successful approaches for speech recognition nowadays [5]. The reusability of such segments, which need to be trained only once and can then be applied to recognize any vocabulary, represents a considerable advantage compared to rigid whole word models.
Although good results are achieved through this flexible technique, little is known about how the size of training data affects the recognition rate. Also, to our knowledge, the flexible vocabulary approach has not sufficiently been compared to other methods by using a similar training corpus and evaluating the obtained models against a common set of data.

The goal of the paper at hand, is to compare different HMM subword model approaches with word models. Since we are not only interested in the final recognition rates after the entire training material has been used, but also on the influence of training data amount, we are also showing for the experiments with phonemes as subword units, how this technique improves performance.

2. EXPERIMENTS

For the intended evaluation, we choose to compare results based on

- phonemes, since those represent the most often used subword units,
- diphones, which are defined as the segment characterizing the transition from the middle of a phoneme to the middle of its next neighbor,
- and whole words, providing us with the necessary reference results.

2.1. Architecture of Models

All three presented units are defined as continuous left-right HMMs and were trained using HTK [4].
As parameter vector, we use 12 Mel Frequency Cepstral Coefficients (MFCCs) and energy, as well as their first and second derivatives. The resulting vector of 36 elements is split up in three streams of 13 stationary coefficients, 13 delta coefficients and 13 acceleration coefficients.
Most Markov Models of phonemes and diphones consist of three emitting and two non-emitting states, following the idea that each of these states would be assigned to a transitional or a stable zone of the subword segment.
The word models adhere to the same principal idea, which results in whole word HMMs with their number of states being approximately three times the number of included phonemes.
The covariance matrices for all three HMM families are assumed to be diagonal, while the number of Gaussian mixtures per state and stream, which represent the observation probability, varies from mostly three (for diphones), over about five (for words), to approximately twenty (for phonemes).

2.2. Training Protocol

The experiments were conducted on parts of the Swiss French PolyPhone and PolyVar databases, described in [2] and [3]. These telephone speech databases fulfill the Polyphone

standard and are recorded in the framework of the SpeechDat project, as a collaboration between IDIAP and Swiss Telecom PTT. Though both databases were recorded on the Swiss ISDN telephone network, different recording platforms were used: PolyPhone calls were recorded on a digital Rhetorex machine, while PolyVar callers spoke to a DialSys4 board based PC, commercializesed by ACSYS.

For training of subword models, a total of 2250 sentences uttered by 231 speakers (9-10 sentences per speaker) were selected from the Swiss French PolyPhone database. The extracted sentences were labeled to the phoneme level semi-automatically. In order to monitor the influence of the amount of available training data upon recognition performance, the material was split into different sets, as described in Table 1, and training of phoneme models was performed on these differently sized pieces of data.
Since for some of the diphones required for the chosen validation vocabulary the number of repetitions in the annotated sentences was quite low – some of them did even not occur at all – it would have not been possible to observe the transition based models' recognition results with respect to training data amount. Thus, the diphone models were trained directly on the entire corpus of phonetically labeled sentences (set F).
For both, phoneme and diphone models, between three and nine Baum-Welsh reestimation cycles were computed before obtaining the best recognition results.

Phoneme training sets	Female speakers	Male speakers	Number of sentences
Set A	26	26	514
Set B	33	33	654
Set C	40	40	793
Set D	52	52	1018
Set E	64	64	1255
Set F	115	116	2240

Table 1: The training sentences for phoneme models were split into different sets for the purpose of evaluating the influence of the amount of training data on recognition performance. Training set A is comprised in set B, which is comprised in set C, and so on.

As reference for the proposed validation, we created two sets of whole word HMMs:

- The (Swiss) French digits from zero to nine, including the words for the hash (#) and star (*) symbols.
  The models were trained on the utterances of 471 speakers from the PolyPhone database, each pronouncing one sequence of six digits (and symbols).
- A vocabulary of 17 application words from the PolyVar database:
  annuler, casino, cinéma, concert, exposition, galerie du manoir, giannada, guide, message, mode d'emploi, louis moret, musée, précédent, quitter, suivant, manifestation.
  Here, the training material consists of recordings of 10 speakers (5 female, 5 male), where each of them repeats all 17 words 30 times.

2.3. Validation and Evaluation Tests

The validation and evaluation of the flexible vocabulary approach against performances of whole word HMMs was attempted for

1. sequences of connected (Swiss) French digits including hash and star, where the utterances of 497 male and female PolyPhone speakers (one sequence of six digits per speaker) other than those from the training corpus were presented to the Viterbi recognizer,
   and for

2. isolated words, where the above enumerated 17 words were pronounced by 103 PolyVar speakers between 1 and 26 times, resulting in a total of 313 isolated words. This test is supposed to give also an idea about the robustness of the evaluated subword units, since PolyPhone and PolyVar calls were recorded on different recording platforms.

Whole word models, diphone models, and phoneme models were tested on the above described corpora, where for the latter class of subword units, performance was evaluated for training sets of different size described in Table 1.

3. RECOGNITION RESULTS

The allover performance obtained with the different approaches for connected digits and isolated words is shown in Table 2. Here, the recognition results of phonemes are achieved with HMMs that were trained on the largest training amount available (set F).

Validation	On connected digits	On 17 isolated words
Whole word HMMs	97.52%	98.13 %
Phoneme HMMs	95.44%	97.10%
Diphone HMMs	84.81%	83.91%

Table 2: Validation of the three approaches through comparison of recognition performances: While phoneme models approach whole word HMM results, diphones are far behind.

The influence of the growing amount of training data on the performance of phoneme models can be seen in Table 3.

Evaluation	On digits (inc. # *)	On 17 words
Set A	86.89%	95.96%
Set B	89.84%	
Set C	90.85%	95.85%
Set D	91.85%	
Set E	93.23%	96.50%
Set F	95.44%	97.10%

Table 3: Evaluation of phoneme HMMs trained on varying amount of data.

4. DISCUSSION

The reported recognition results clearly validate the flexible vocabulary approach. Even though whole word models still outperform the subword unit approaches, we obtain with phoneme HMMs comparable results. The diphone models, however, appear not very successfull.

4.1. Diphones as Units

Although the transitions between the stable phoneme zones comprise much acoustic and articulatory information, the here presented results could be classified as dissapointing, at first. However, there are a few reasons, why these subword segments perform rather poorly.

The first major problem is the lack of training data. Almost 20% of the approximately 300 diphones which are contained in the vocabularies used for validation, occured less than 10 times in our training data. About 60% of the required transition segments were represented less than 100 times. This obvious handycap forced us even to use simpler HMM topologies than for phonemes, since the models could otherwise hardly converge during training.

Furthermore, the grammar for the Viterbi recognizer proved to be difficult to elaborate for the connected digits experiment. We encountered the problem of defining beginnings and endings of the different digits. This interfered also with the problem of short words: the French digit "un" (one), for instance, is pronounced as one of two alternative phonemes and was recognized only 27.56% of the time. But also rather long vocabulary items, like "galerie du manoir", were less recognized (37.58%) than the average.

4.2. Phonemes as Units

Unlike for diphones, the performance of phoneme models is quite encouraging. The recognition rates of these elements approach those of whole word models to a difference of only 1%.
This result becomes even more interesting when we consider, that training and evaluation were performed on utterances recorded on different platforms.
Furthermore, the phoneme models were trained without regard of context.

The last two facts lead us to the conclusion, that simple context independent phoneme models might be sufficient for a number of applications, which require speaker independent recognition of isolated or connected words.
Also, the task of creating a grammar for the Viterbi recognizer is unproblematic, which is a further advantage.

4.3. Influence of training data

The influence of the amount of available training material on recognition rates obtained with phoneme models is not surprising. Nevertheless, one might wonder, why the improvements of phoneme based performance are more visible for the connected digits task, than for the set of 17 application words.

The explanation herefore lies in the phonetic decomposition of the two vocabularies: the digits make use of two phonemes, which are rather under-represented in the sentences used for training, especially in the first two data sets. In the sametime, these two phonemes do not appear in the second test vocabulary with the 17 application words.
Since one of these poorly represented phonemes is "9~", the more common of two possible pronunciations of the French digit "un" (one), the insufficiently trained HMM provokes a noticeable recognition error for this digit, which lowers the overall performance considerably, as shown in Table 4.

Training on	Training samples of 9~	Training samples of e~	Recognition of "un" (one)	Degradation of total performance
Set A	131	168	36.47%	5.43%
Set B	163	210	52.94%	4.02%
Set C	194	240	60.39%	3.39%
Set D	246	331	65.49%	2.25%
Set E	286	432	77.95%	1.88%
Set F	541	906	87.06%	1.11%

Table 4: Recognition rate of the French digit "un" (one) in the connected digits sequences and contribution to total error for the digits vocabulary.

Another result worth mentioning can be observed on two similar words of the 17 application words: "guide" (phonetic: "gid" with an optional appended "@") and "quitter" (phonetic: "kite"). Here, the first phoneme "g" of the former word is also rather less represented in our training material, which causes the recognizer to mistake "guide" for "quitter" quite often, as shown in Table 5.

Training on	Samples of phoneme "g"	Recognition of "guide"	Confusion with "quitter"
Set A	123	84.95%	12.22%
Set C	184	80.25%	16.67%
Set E	301	85.27%	11.60%
Set F	603	86.52%	10.97%

Table 5: Insufficiently trained phoneme "g" causes confusion between partially similar sounding words.

5. CONCLUSION

The results obtained through this work show, that good subword models can be almost as performant as whole word HMMs, where the gap between recognition rates of the considered approaches is as small as 1%.
The here presented findings validate the idea of using flexible subword units for recognizing speaker independent isolated words, as well as connected words.

The amount of training data required for adequate phoneme HMMs depends on the length and phonemic similarity of the words constituting the envisaged vocabulary.

Large amount of necessary training data and the non trivial task of creating a sensible grammar for connected word (digits) recognition, make diphones as subword segments less attractive.

Context and speaker independent phoneme HMMs show good performances on isolated and connected words recognition tasks and demonstrate also a certain robustness.

6. ACKNOWLEDGMENTS

The recording of the mentioned Swiss French PolyPhone and PolyVar databases was partially funded by the Federal Office for Education and Science (OFES) in Switzerland and by Swiss Telecom-PTT.

7. REFERENCES

1. G Chollet. New advances and trends in speech recognition and coding., chapter Evaluation of ASR systems, algorithms and databases. Nato ASI, Bubion, 1994.
2. G Chollet, JL Cochard, A Constantinescu, P Langlais, and R Van Kommer. Swiss French Polyphone and Polyvar : Telephone Speech Databases to Model Inter- and Intra-Speaker Variability. Technical report, IDIAP, 1996.
3. A Constantinescu and G Chollet. Swiss polyphone and polyvar: Building databases for speech recognition and speaker verification. 3rd Slovenian-German and 2nd SDRV Workshop, Speech and Image Understanding, Ljubljana, April 1996.
4. Entropic Research Laboratory. Using HTK to design a Speaker independent connected digit recognition system, 1994.
5. Hwang MY Huang XD, Hon HW and Lee KF. A comparative study of discrete, semi-continuous and continuous hidden markov models. Computer Speech and Language, 7(4), 1993.
6. L Rabiner and Biing-Hwang Juang. Fundamentals of Speech Recognition. Prentice-Hall signal processing series, 1993.
7. Michael D. Riley. Speech Time-Frequency Representations. Kluwer Academic Publishers, 1989.