HYBRID HMM/ANN SYSTEMS FOR TRAINING INDEPENDENT TASKS: EXPERIMENTS ON PHONEBOOK AND RELATED IMPROVEMENTS

Stéphane Dupont[1], *Hervé Bourlard*[2,3], *Olivier Deroo, Vincent Fontaine & Jean-Marc Boite*

Faculté Polytechnique de Mons — TCTS
31, Bld. Dolez
B-7000 Mons, Belgium
Email: dupont,bourlard,deroo,fontaine,boite@tcts.fpms.ac.be

ABSTRACT

In this paper, we evaluate multi-Gaussian HMM systems and hybrid HMM/ANN systems in the framework of task independent training for small size (75 words) and medium size (600 words) vocabularies. To do this, we use the PHONEBOOK database [6] which is particularly well suited to this kind of experiments since (1) it is a very large telephone database and (2) the size and content of the test vocabulary is very flexible. For each system, different HMM topologies are compared to test the influence of state tying (with a number of parameters approximately kept constant) on the recognition performance. Two lexica (Phonebook and CMU) are also compared and it is shown that the CMU lexicon is leading to significantly better performance. Finally, it is shown that with a quite simple system and a few adaptations to the basic HMM/ANN scheme, recognition performance of 98.5% and 94.7% can easily be achieved, respectively on a lexicon of 75 and 600 words (isolated words, telephone speech and lexicon words not present in the training data).

1. INTRODUCTION

Task independent training remains (among many others!) an important issue in current automatic speech recognition (ASR) systems. It is indeed well know that ASR performance is always significantly lower for lexicon words that were not observed in the training data. Also, given that most state-of-the-art ASR systems use context dependent phone models, it is not clear how to generalize them in the case of new words [3]).

In this paper, hybrid HMM/ANN (Hidden Markov Models - Artificial Neural Networks) and multi-Gaussian HMM systems are tested in the framework of

- Context independent and "some kind of" generalized context dependent phone models.
- Training independent tasks, for small (75 words) and medium (600 words) size vocabularies.

In this framework, the PHONEBOOK [6] database has been used. Since, to our knowledge, we are one of the first ones to report results on this task and since no formal training and test sets have been defined yet, we start by clearly defining below how we have split up the data into training, crossvalidation and test sets.

2. DATABASE

All the experiments reported in this paper have been carried out on the PHONEBOOK [6] database. This is a phonetically-rich isolated word telephone-speech database. PHONEBOOK consists of more than 92,000 utterances and almost 8,000 different words, with an average of 11 talkers for each word. Each speaker of a demographically-representative set of over 1,300 native speakers of American English made a single telephone call and read 75 words.

The database contains 106 word lists, each composed of 75 or 76 words that have been pronounced by a few (typically around 11) speakers. The speakers and words are different for each word list. The word lists are labeled as $l_1 l_2$ with

$$l_1 \in \{a, b, c, d, e\}$$

and

$$l_2 \in \{a, b, c, d, e, f, \ldots, x, y, z\}$$

except if l_1=e, in which case l_2 is then equal to a or b only. There are thus 106 word lists. The database being very large (totaling 23 hours of speech, μ-law coded), we defined two training sets, one cross-validation set and one test set as follows:

- *a "small" training set totaling approximately 5 hours of speech*:
  all $*a$, $*h$, $*m$, $*q$, and $*t$ word lists, i.e., 21 word lists.

[1] Supported by a F.R.I.A. grant (Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture).

[2] Also affiliated with Intl. Computer Science Institute, Berkeley, CA.

[3] Now with IDIAP, Switzerland.

- *the "full" training set totaling approximately 21 hours of isolated words*:
  all word lists except the ones present in the cross-validation set and in the test set.
- *the cross-validation set*:
  all $*o$ and $*y$ word lists (8 word lists).
- *the test set*:
  all $*d$ and $*r$ word lists, i.e., 8 word lists. Since the lexicon is different in each of these 8 word lists, we then have the choice to recognize the 8 word lists as a whole (yielding a lexicon of 600 words) or to recognize each word list independently with a lexicon of about 75 words. In the second case, the recognition rate will be the (unweighted) average over the 8 recognition rates.

So far, only the "small" training set has been used.

Two dictionaries have been used to generate two independent sets of phonetic transcriptions of the training words as well as the lexicon words of the test sets. The first dictionary is the one released with PHONEBOOK and contains the phonetic transcriptions of the PHONEBOOK words according to a 42-phoneme inventory. The second dictionary is the 110,000-word CMU 0.4 dictionary using 39 phonemes (a subset of the TIMIT phonemes). Some of the PHONEBOOK words that were not present in CMU 0.4 have been transcribed manually.

3. ACOUSTIC FEATURES

Two sets of acoustic features have been used: the log-RASTA-PLP cepstral features [2] and the lpc-cepstral features with cepstral mean subtraction (CMS). These features have been chosen for their robustness against channel and speaker characteristics. These parameters were computed every 10 ms on 30 ms analysis windows. The order of the lpc analysis was set to 10.

The feature set for our hybrid HMM/ANN systems was based on a 26 dimensional vector composed of the *cepstral* parameters (log-RASTA-PLP or lpc-cepstral parameters with cepstral mean subtraction), the $\Delta cepstral$ parameters, the $\Delta energy$ and the $\Delta\Delta energy$. Nine frames of contextual information was used at the input of the ANNs, leading to 234 inputs (9 frames being known as yielding the best recognition performance). Continuous density systems used an extended vector of 38 components containing the $\Delta\Delta cepstral$ parameters.

4. RECOGNIZERS

In the following experiments, hybrid HMM/ANN systems [1] as well as continuous observation densities

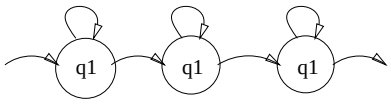

Figure 1. Context independent phoneme model with 3 tied states.

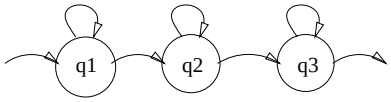

Figure 2. Context independent phoneme model with 3 independent states.

HMMs (Gaussian mixtures) [4] have been used and compared.

Three kind of left-to-right phone models have been tested:

1. *Model 1*: 1-state phone models with a minimum duration constraint of 3, as presented in Figure 1.
2. *Model 2*: 3-state phone models (minimum duration of 3), as shown in Figure 2.
3. *Model 3*: 3-state phone models with tied states for the 1st and the 3rd states (Fig. 3). These states are tied across different phones and possibly across all the phones. They represent the inter-phone transitions. In the experiments reported here, tying was done across the phones according to their broad phonetic class. This lead to 9 transition states towards a phone and 9 transition states from a phone.

It was initially expected that *Model 3* could lead to better recognition performance in the case of training independent tasks. Indeed, the training data always contains the phonemic units in a limited number of left and right contexts. In standard recognizers (and standard reference tasks), it is possible to make use of this contextual information to improve performance, e.g., by using context dependent phone model. Even when using context independent phone models (which is often the case with hybrid HMM/ANN systems that are however known to yield comparable – although still somewhat lower – performance compared to context-dependent models), the phone models will implicitly capture some contextual information. However, if the application (or test) vocabulary is different from the

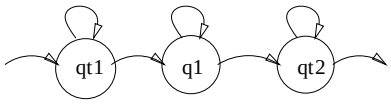

Figure 3. Context independent phoneme model with 1 state q_1 and 2 transition states q_{t1} and q_{t2}. These two states are tied across several phones.

Model type	Parameters	log-RASTA	CMS
Model 1	166k	7.7%	5.1%
Model 2	166k	5.3%	4.1%
Model 3	166k	5.4%	4.3%

Table 1. Error rates on isolated word recognition (75 lexicon words never seen in the training data) with hybrid HMM/ANN systems and either log-RASTA-PLP feature set or CMS feature set. The 3 kinds of phone models proposed in Section 3 were tested.

Model type	Mixtures /state	Parameters	CMS
Model 1	48	162k	7.1%
Model 2	16	162k	5.0%
Model 3	34	162k	6.2%

Table 2. Error rates on isolated word recognition (75 lexicon words never seen in the training data) with continuous densities HMMs and the extended CMS feature set. The 3 kinds of phone models proposed in Section 3 were tested.

training vocabulary, using that phonemic contextual information could result in a loss of performance since the trained models are no longer really appropriate. *Model 3* attempts to limit this effect by tying the distributions of the first and last states of *Model 2* across several phonemes. In this way, it can be expected that all major contextual effects will be captured by those tied "transition" states (yielding the same contribution for all phonemes during recognition) while the middle state will focus on the actual "steady-state" section of each phonemic segment. As opposed to SPAM [5], which was attempting to focus on transitions only, *Model 3* could be referred to as "anti-spam" since it aims at focusing on steady-state segments only. Although Model 3 was indeed initally yielding better performance for small training data sets, this conclusion was actually not confirmed for the larger experiments reported here.

In all cases, training was done by embedded Viterbi.

5. RECOGNITION EXPERIMENTS

Experiments have been performed to compare the three kinds of models. We used the phonetic transcriptions released with PHONEBOOK and trained systems on both log-RASTA-PLP parameters and CMS parameters. Tables 1 and 2 summarize the results achieved for hybrid HMM/ANN and multi-Gaussian systems on a 75 isolated words recognition task.

The multi-Gaussian systems used diagonal covariance matrices, and the number of Gaussians per state

Model type	Parameters	log-RASTA	CMS
Model 1	166k	2.4%	1.5%
Model 2	166k	2.9%	1.8%
Model 3	166k	2.6%	2.1%

Table 3. Error rates on isolated word recognition (75 lexicon words never seen in the training data) with hybrid HMM/ANN systems and either log-RASTA-PLP feature set or CMS feature set. The 3 kind of phone models proposed in Section 3 were tested (with minimum duration modeling). Transcriptions from the CMU 0.4 dictionary were used.

was chosen to keep the number of parameters across the different experiments approximately constant.

Similar experiments were also performed with transcriptions based on the CMU 0.4 lexicon (in place of the PHONEBOOK lexicon) with a significant performance improvement. Since they were better in the initial experiments, we only tested HMM/ANN systems. Just changing the lexicon, the error rate for *Model 1* went down from the 7.7% in Table 1 to 3.6%. Consequently, all the following experiments were done with the CMU 0.4 dictionary. As a second step to reduce the error rate, we used minimum duration modeling. This minimum duration was half of the mean duration of the considered state (computed on a forced Viterbi alignment of the training data). From 3.6% error rate, we went down to 2.4%. Complete results for the three model types and with minimum duration modeling are reported in Tables 3 (75 words) and 4 (600 words). As shown in Table 3, the best system was yielding 1.5% error rate on the 75 word test vocabulary. With the same system, we also achieved 5.3% error rate on the 600 word lexicon (Table 4).

It was expected that *Model 3* could yield better performance than the other models. The general idea was to capture all major contextual (inter-phoneme) effects with a few transition states, while still focusing on the less-coarticulated middle part of each phonemic segment. Clearly, the experiments were inconclusive in this respect. The 3 models are indeed yielding comparable performance (with a slight preference for *Model 1*) in the case of state-of-the-art systems. This can however be explained by the large amount of data that was used to train our systems. Phonemes are presented in a sufficient number of left and right contexts, yielding to efficient and robust classical context independent models, even when the test vocabulary is different from the training vocabulary.

Model type	Parameters	log-RASTA	CMS
Model 1	166k	7.8%	5.3%
Model 2	166k	8.7%	5.6%
Model 3	166k	7.6%	5.9%

Table 4. Error rates on isolated word recognition (600 lexicon words never seen in the training data) with hybrid HMM/ANN systems and either log-RASTA-PLP feature set or CMS feature set. The 3 kind of phone models proposed in Section 3 were tested (with minimum duration modeling). Transcriptions from the CMU 0.4 dictionary were used.

6. CONCLUSIONS

In this paper, training independent isolated word recognition was investigated in the framework of both multi-Gaussian systems and discriminant HMM/ANN systems. Good performance was achieved on small and medium vocabulary tasks with quite classical systems. Different HMM topologies were compared to test the influence of different level of state tying. Confirming previous experiments (from us and others), the best results were obtained with single state HMM/ANN phoneme models with minimum duration modeling. All the experiment were done with the STRUT (Speech Training and Recognition Unified Toolkit) software [7].

ACKNOWLEDGMENTS

We thank the European Community for their support in this work (SPRACH Long Term Research Project 20077).

REFERENCES

[1] Bourlard, H. and Morgan, N., *Connectionist Speech Recognition – A Hybrid Approach*, Kluwer Academic Publishers, 1994.

[2] Hermansky, H. and Morgan, N., "RASTA processing of speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4 pp. 578-589, 1994.

[3] Hon, H., and Lee, K., "On vocabulary-independent speech modeling" *Proceedings of ICASSP'90*, pp. 725-728, 1990.

[4] Juang, B.H., Levinson, S.E., and Sondhi, M.M., "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Trans. Information Theory*, IT-32 (2): 307-309, March 1986.

[5] Morgan, N., Wu, S., and Bourlard, H., "Digit recognition with stochastic perceptual speech models" *Proceedings of EUROSPEECH'95*, Madrid, 1995.

[6] Pitrelli, J., Fong, C., Wong, S., Spitz, J., and Leung, H., "Phonebook: A Phonetically-Rich Isolated-Word Telephone-Speech Database" *Proceedings of ICASSP'95*, Detroit, Michigan, 1995.

[7] "STRUT Home Page " `http://tcts.fpms.ac.be/speech/strut.html`.