Context dependent modelling approaches for hybrid speech recognizers

Alberto Abad[1], *Thomas Pellegrini*[1], *Isabel Trancoso*[1,2] *and João Neto*[1,2]

[1]INESC-ID Lisboa, Portugal
[2]Instituto Superior Técnico, Lisboa, Portugal

alberto.abad@inesc-id.pt

Abstract

Speech recognition based on connectionist approaches is one of the most successful alternatives to widespread Gaussian systems. One of the main claims against hybrid recognizers is the increased complexity for context-dependent phone modeling, which is a key aspect in medium to large size vocabulary tasks. In this paper, we investigate the use of context-dependent triphone models in a connectionist speech recognizer. Thus, most common triphone state clustering procedures for Gaussian models are compared and applied to our hybrid recognizer. The developed systems with clustered context-dependent triphones show above 20% relative word error rate reduction compared to a baseline hybrid system in two selected WSJ evaluation test sets. Additionally, the recent porting efforts of the proposed context modelling approaches to a LVCSR system for English Broadcast News transcription are reported.

Index Terms: speech recognition, context modeling, connectionist system

1. Introduction

Hidden Markov Models of Gaussian mixtures (HMM/GMM) is doubtless the most widely accepted framework for automatic speech recognition (ASR). Alternatively, Artificial Neural Networks (ANN) have also been proposed [1], but despite their high discrimination ability in short-time classification tasks, they have proved inefficient when dealing with long-term speech segments. With the goal of solving this problem, one of the most successful alternatives to HMM/GMM was later proposed, commonly known as hybrid ANN/HMM or connectionist paradigm [2]. In general, hybrid architectures seek to integrate the ANN ability for estimation of Bayesian posterior probabilities into a classical HMM structure that allows modeling the long-term speech evolution.

The main advantage of hybrid approaches is that classification networks are usually considered better pattern classifiers than Gaussian mixtures approaches. Additionally, an appealing characteristic of the hybrid systems is that they are very flexible in terms of merging multiple input streams. One of the most significant limitations, however, is related with the lack of flexibility and increased difficulty when context-dependent phone modeling is desired. Some approaches to phonetic context training in ANN are reported in [3, 4] based on factorization of posterior probabilities. In [5] triphone modeling is introduced by interpreting the probability outputs of the neural network as the codebook of a tied-mixture system.

With the aim of improving our connectionist ASR system (AUDIMUS [6]) by means of a better modeling of context dependencies, we proposed combining multiple state subphoneme recognition units with a restricted set of diphones [7]. In the present work, we investigate the use of context-dependent triphone neural networks. Similarly to [8], we interpret the probability outputs of the neural network as clustered triphone models. However, we do not rely on an HMM/GMM recognizer for forced alignment generation of context-dependent triphone states and for clustering computation. Alternatively, we first propose generating state-level triphone alignments by expanding the multiple-state monophone alignments described in [7]. Then, we focus on triphone clustering approaches. The clustering step is crucial since each triphone clustered state will be represented as a neural network output in the connectionist system. Thus, several experiments with both data-driven and decision tree-based approaches and with different number of final clusters are presented. The developed triphone state-clustered systems are compared with the monophone connectionist system. Recent experiments with Broadcast News are also reported.

2. Corpora and task description

The experiments reported in this work refer to two very different tasks. The earlier experiments were done with the Wall Street Journal (WSJ) read speech corpus [9]. Only the SI-84 training material from WSJ0 was used, resulting in approximately 15 hours. The November 1992 ARPA WSJ evaluation corpora (*Nov92*) containing 330 sentences from 8 speakers is used as development set. The *si_dt_s6* data from WSJ1 (202 sentences from 8 speakers) and the *si_dt_05.odd* subset of WSJ1 (248 sentences from 10 speakers) defined in [10] are used as evaluation sets. The assessment was done using the WSJ 5K non-verbalized 5k closed vocabulary set and the WSJ standard 5K non-verbalized closed bigram language model.

The later experiments were done in a broadcast news domain. The speech material for acoustic model training was the 1996 (LDC97S44) and 1997 (LDC98S71) English BN Speech corpus (HUB-4), comprised of respectively 73 and 67 hours of manually transcribed speech, coming from several television and radio networks. The corresponding orthographic transcriptions contain respectively 850k and 790k words. An additional corpus of transcriptions, named CSR'96 (LDC1998T31), totalling 149M words, was also used for language modeling, together with two newspaper and newswire text corpus: the North American News Text Corpus (LDC1995T21), with 505M words, and the LDC1998T30 corpus, with 462M words. Four official NIST data sets were used for testing: LDC2000S86, LDC2002S11, LDC2000S88 and LDC2007S10.

3. Systems description

3.1. The baseline AUDIMUS hybrid speech recognizer

AUDIMUS is based on the hybrid ANN/HMM paradigm for speech recognition [2]. This kind of recognizers are generally composed by a phoneme classification network, particularly a MultiLayer Perceptron (MLP), that estimates the posterior probabilities of the different phonemes for a given input speech frame (and its context). These posterior probabilities are associated to the single state of context independent phoneme HMMs. Concretely, the system combines the outputs of three MLPs trained with Perceptual Linear Prediction features (13 static + first derivative), log-RelAtive SpecTrAl features (13 static + first derivative) and Modulation SpectroGram features (28 static) [6]. The decoder is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition, that maps observation distributions to words.

3.2. Single-state monophone system (SS)

The TIMIT database with manually annotated phonetic alignments is first used to train MLPs that are then used for the generation of frame-to-phone alignments of the training data using word-level transcriptions.

The process for the estimation of the final phone classification networks consists of several iterations of re-alignment and re-training until a stable phone classification performance is obtained in the development test set (*Nov92*).

In this work, the trained MLPs are composed of an input layer with a context window of 7 frames, 2-hidden layers of 700 weights each one and 40 units output layer (40 phonemes including the silence pattern). Although the resulting networks are relatively large for the available amount of data (∼650K parameters and ∼7.5 patterns per weight), this size has been experimentally found to be the most adequate.

3.3. Multiple-state monophone system (MS)

The extension of the baseline ANN/HMM speech recognizer based on single state HMM models to multiple state models was described in [7]. The underlying idea derives from the characteristics of the phone production process. Each phone is usually considered to be constituted by three regions: an initial transitional region, a second central steady region known as phone nucleus, and a final transitional region. Thus, it is expected that modeling each one of these portions separately will produce an improvement of the acoustic phone modeling and consequently and improvement in the recognition performance. An initial "blind" alignment derived from the single-state alignment is used for boot-strapping the MS neural network. Then, multiple iterations of re-alignment and re-training follow, like in the case of the baseline system. The resulting sub-phoneme classification networks have 118 outputs (silence is kept as a single state monophone) instead of the 40 outputs of the baseline system. The total number of parameters of each stream network is approximately 700K.

3.4. Acoustical modeling of phone transitions (PT)

The previous multi-state model was also extended to incorporate modeling the most significant/frequent transition units as additional sub-phone recognition units. The three-state model of left and right general class context-dependent and nucleus units is still necessary to give coverage to all the data. In order to generate an initial alignment with phone transition modeling, the sub-phoneme forced alignment provided by the multiple state hybrid system is transformed. Context-right and context-left units that had an equivalent selected word-internal transition unit (frequent enough) are replaced by the corresponding context-dependent model. This alignment is used to train initial networks. Then, iterations of re-training and re-alignment are done until phone classification performance converges in the development set.

3.5. Results

Table 1 shows the performance results of the three above mentioned reference systems: the single-state and multiple-state ANN/HMM multiple stream systems, and the one including modeling of phone transitions.

System	Nov92	si_dt_s6	si_dt_05.odd	#Params
SS	9.77	13.13	14.37	650K (x3)
MS	8.74	12.59	13.56	700K (x3)
PT-20% (14)	8.52	11.53	12.52	710K (x3)
PT-40% (40)	8.66	11.23	11.91	730K (x3)
PT-50% (61)	7.79	10.27	11.88	740K (x3)
PT-60% (91)	8.28	10.09	11.64	760K (x3)
PT-70% (137)	7.98	9.22	11.22	800K (x3)
PT-80% (203)	7.57	9.61	10.95	840K (x3)

Table 1: *WER results of the baseline single-state (SS), multiple-state (MS), and phone transition (PT) ANN/HMM systems.*

A considerable improvement was achieved thanks to multiple-state modeling with respect to the single-state monophone system. However, it is worth noticing that very well-trained SS networks are preferable but not crucial, since they are only used as a previous step for generating multiple-state alignments. In some cases, a better initial monophone state-level alignment may affect the speed convergence of the next stages, but it has been generally observed an insignificant influence in the speech recognition performance of the resulting context-dependent systems.

The last six lines of the Table show the results achieved for different coverage rates of transition units instances appearing in the training data. The phone transition approach outperforms both the single-state and the multiple state hybrid system independently on the number of transition units selected. The best system is the one with 80% of diphones coverage. In this case, compared to the baseline SS hybrid system, a 26.8% and 23.8% relative WER reduction is achieved for the *si_dt_s6* and *si_dt_05.odd* evaluation test sets respectively.

4. Triphone modeling

The use of context-dependent models in both HMM/GMM and ANN/HMM frameworks becomes a problem due to the huge number of possible triphones in medium to large vocabulary recognition tasks. As a consequence, Gaussian systems result in an increased number of models, while hybrid systems require neural networks with large output layers to be trained. In both cases, the enormous increase in the number of parameters leads to sparse data problems. In practice, triphone state clustering methods are commonly applied in order to reduce the total amount of physical context dependent models in Gaussian systems. Thus, sparcity problems can be effectively solved leading to more robust parameter estimation. In this work, we intend to investigate and compare the same clustering methods commonly used in Gaussian systems applied to connectionist recognizers. Concretely, we apply both data-driven (DD) and decision tree (DT) clustering approaches to reduce the number of state triphones to different desired amounts.

In order to obtain the necessary triphone state-level alignment for triphone clustering application, the monophone state-level forced alignments obtained with the well-trained multiple-state connectionist system of previous section are expanded to word-internal triphones. The main reason for considering only word-internal is that our decoder system does not allow cross-word dependency decoding strategies.

The initial number of different triphones present in the training data is 7043, which corresponds to 21129 states (3 states per model). Triphones corresponding to the central *silence* model are not included in the 7043 triphones, neither are used in the clustering procedure. Silence was always modeled in our experiments as a single output of the neural networks.

4.1. Data-driven clustering

Data-driven triphone state clustering allows to choose which states should be tied according to some distance between any two states. Similar to the clustering procedure provided in the HTK toolbox [11], a hierarchical bottom-up clustering algorithm has been developed, in order to experiment various distance metrics, and to control the clustering procedure better. Initially, each state is a single cluster characterized by a vector formed by the mean and variance of the frames aligned with that state. Pairs of the closest clusters in terms of a predefined distance are merged until some stop criterion is met. After each merge, distances to the newly formed cluster are updated with the so-called complete link method: the updated distance is the farthest distance between any two patterns in the two clusters.

The clustering procedure is done in two steps. First, the states with an occupation count smaller than a fixed threshold are used for clustering. The occupation count is the number of frames attributed to a state, and can be seen as a measure of the quantity of training data per model. The threshold value is subject to some tuning, and after some experimentation, the minimum occupation was fixed to 200. In a second step, the clustering is free to chose the state pairs according to the normal distance criterion, and merges states until the number of desired units is reached.

Notice that state pairs clustering is only allowed between triphones of the same phone class and *position* state, that is, *left*, *center* or *right* state of the three-state model. In other words, the clustering procedure with the minimum number of clusters is the one that corresponds to multiple state monophone modeling of previous section with 118 models (117 states + 1 *silence*).

Various distance metrics have been tested: Euclidean, Weighted Gaussian (as defined in page 269 of [11]), Mahalanobis and Bhattacharryya distances. To quantitatively compare the different clustering models, the cophenetic correlation coefficient (CCC) was evaluated, comparing the original distances between the different patterns, to their distances after clustering. The closer the coefficient to 1, the more relevantly the clustering represents the data structure [12]. The Euclidean distance gave the biggest CCC, with a 0.57 value, thus only the results achieved with this distance are reported henceforth.

4.2. Decision tree-based clustering

Decision-tree based clustering is an alternative approach for triphone clustering. It is usually preferable to DD clustering, since it offers a solution for dealing with triphones for which there are no examples in the training data.

In this work, we have used the HHEd tool of the HTK toolkit for DT clustering. HHed builds top-down decision trees by sequentially finding the binary questions that provide the best split given a set of questions about the left and right contexts of each triphone. In order to use this tool, we transformed the mean and variances of each triphone state computed previously to the HTK format for the definition of single mixture Gaussian models. It was also necessary to convert the occupation statistics of each state to the adequate format. Then, decision-tree clustering was applied to reach a fixed number of final clusters. Like in DD clustering, the creation of clusters with very little associated training data (less than 200 frames) was avoided.

The unseen triphones present in the test lexicon were synthesized with the trees constructed by the HHed tool. Detailed information of the HHed tool and the tree-based clustering procedure method applied can be found in [11].

4.3. Data-driven vs Decision tree-based clustering results

Triphone neural networks have been trained for a different number of clustered state triphones for both DD and DT based clustering. The size of the input and hidden layers has been kept identical to previous monophone networks and only the output layer has been modified according to the number of triphone states. The process for training the neural networks consisted of first transforming the triphone-labeled state-level alignments to clustered triphone states, according to the results of the clustering procedure. Then, once the neural networks were trained, several iterations of re-alignment and re-training were realized until a stable phone classification rate was achieved.

Table 2 shows the recognition results of the triphone connectionist systems. Notice that the number of softmax outputs of the neural network corresponds to the number of clustered states plus one output for the *silence*, since it was not included in the clustering process. For instance, *DD200* stands for neural networks with 200 outputs: 199 data-driven clustered triphone states and 1 *silence* output.

System	Nov92	si_dt_s6	si_dt_05.odd	#Params
DD200	7.85	9.94	12.62	760K (x3)
DD300	7.77	10.99	11.35	830K (x3)
DD500	7.64	10.42	10.81	970K (x3)
DT200	7.75	11.36	12.08	760K (x3)
DT300	7.27	11.11	11.35	830K (x3)
DT500	7.38	11.20	11.95	970K (x3)

Table 2: *WER results of hybrid ANN/HMM systems for different number of context-dependent triphones with both data-driven (DD) and decision tree-based (DT) clustering.*

Results show that it is possible to train neural networks modeling clustered triphone states. All the developed recognizers, independently of the number of clusters or the clustering method considered, clearly outperform the baseline monophone hybrid speech recognizer (both single-state and multiple-state).

It is not clear which is the most appropriate number of clusters. The performance is similar independently of the size of the output layer. It would be interesting to evaluate larger sizes in the future. For instance, Gaussian systems for the same task reported in [7] model 3701 and 3820 tied-states for the *wint* and *xword* cases. However, a large number of network outputs like that would likely lead to undertraining problems and large training times. In that case, it would probably be necessary to carefully design the networks dimensionality (throughout this work the size of the input and hidden layers have been kept fixed) and to increase the amount of training data.

There is not a clear advantage for any one of the clustering methods investigated, but it seems that DD performs slightly better. This result could be considered surprising since DT is supposed to implement a more robust method for synthesizing the unseen triphones. Anyway, a detailed analysis of the estimated clusters by each method has not been done yet, but it is likely that the clusters do not differ too much which could be partially due to the small number of desired units.

In contrast to what we initially expected, clustered triphone models does not show performance improvements compared to our initial approach of modeling phone transitions. However, we expect that the clustering triphone approach could obtain more signifcant improvements than the use of a restricted set of transition units when modeling larger training data sets.

5. Application to LVCSR

The proposed MS and PT approaches have been ported to our English LVCSR system for BN transcription. The HUB-4 1996 and 1997 data sets were used to train networks modeling a set of 39 multiple-state monophones plus two single-state non-speech models (one for silence and one for breath) and 336 phone transition units (chosen to cover more than 90% of all the transition units present in the training data). The number of final recognition units (output layer size) totalizes 455 units. For language modeling, we built one LM per source, including speech transcripts and written text sources. Nine LM were linearly interpolated with optimization of the weights on a subset of the HUB-4 1997 training corpus used as development corpus. The final interpolated language model is a 4-gram LM, with Kneser-Ney modified smoothing, comprised of 64k words (or 1-grams), 12 M 2-grams, 5.8M 3-grams and 4.5M 4-grams.

The 64k words vocabulary consists of all the words contained in the HUB-4 training set plus the most frequent words in the broadcast news texts and Newspapers texts. The pronunciations were extracted from the public domain lexicon provided by CMU. For words not included in this lexicon, a rule-based grapheme-to-phone conversion system was used. The multiple-pronunciation lexicon included 70k entries. Table 3 shows the performances achieved on four official NIST test sets.

The new system with context-dependent model units outperforms the monophone system baseline by more than 20% relative performance gains. The *eval98* set shows the biggest improvement, with a 24.1% relative reduction in WER. A slightly worse performance is achieved on the *eval99* test set, maybe due to the lack of contemporary training data.

System	eval97	eval98	eval99	eval03
SS	27.5	26.9	29.4	26.5
MS+PT	22.0	20.4	23.3	20.6

Table 3: Word Error Rates (WERs) achieved on four NIST evaluation test sets, with the monophone baseline and the proposed method of transition phone modeling.

6. Conclusions

It is a well-known fact that context-dependent triphone modeling in Gaussian-based systems allows drastic performance improvements with respect to monophone recognizers in medium to large size vocabulary tasks. In this paper, it was shown that hybrid ANN/HMM speech recognition systems can also benefit from triphone modeling. The investigated connectionist systems with triphone networks clearly outperform conventional monophone recognizers. In order to avoid typical undertraining problems, both data-driven and decision tree-based approaches for triphone state clustering have proved effective.

The results obtained were affected by the reduced number of triphone states considered and also by the fact that some context is already accounted for by the input layer of neural networks. In this way, future work will be focused on the increase of the number of triphones, incorporating cross-word triphones, modifying the networks dimensionalities and considering large vocabulary tasks with additional training data. Indeed, the initial porting efforts of the context modelling approaches for hybrid systems in a LVCSR system for BN transcription have been reported in this paper showing encouraging results.

7. Acknowledgements

This work was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds, and also through project PT-STAR.

8. References

[1] Lippmann, R. P., “Review of neural networks for speech recognition”, Neural Computation, 1(1):1–38, 1990.

[2] Morgan, N. and Bourlad, H., “An introduction to hybrid HMM/connectionist continuous speech recognition”, IEEE Signal Processing Magazine, 12(3):25–42, 1995.

[3] Bourlard, H., Morgan, N., Wooters, C. and Renals, S., “CDNN: A Context Dependent Neural Network For Continuous Speech Recognition”, In Proc. ICASSP’92, II:349–352, 1992.

[4] Franco, H., Cohen, M., Morgan, N., Rumelhart, D. and Abrash, V., “Context-dependent Connectionist Probability Estimation in a Hybrid Hidden Markov Model-Neural Net Speech Recognition System”, Computer Speech and Language, 8(3):211–222, 1994.

[5] Rottland, J. and Rigoll, G., “Tied posteriors: an approach for effective introduction of context dependency in hybrid NN/HMM LVCSR”, In Proc. ICASSP’00, III:1241–1244, 2000.

[6] Meinedo, H., Caseiro, D., Neto, J. and Trancoso, I., “AUDIMUS.media: A Broadcast News speech recognition system for the European Portuguese language”, In Proc. of Int. Conf. of Computational Processing of Portuguese Language (PROPOR), 2003.

[7] Abad, A. and Neto, J., “Incorporating acoustical modeling of phone transitions in an hybrid ANN/HMM speech recognizer”, In Proc. Interspeech’08, 2394–2397, 2008.

[8] Pavelka, T. and Kral, P., “Neural network acoustic model with decision tree clustered triphones”, In Proc. of IEEE Workshop on Machine Learning for Signal Processing (MLSP’08), 216–220, 2008.

[9] Paul, D. and Baker, J. M., “The Design for the Wall Street Journal-based CSR Corpus”, in DARPA Speech and Natural Language Workshop, 1992.

[10] Woodland, P.C., Odell, J.J., Valtchev, V. and Young, S.J., “Large Vocabulary Continuous Speech Recognition Using HTK”, In Proc. ICASSP’94, II:125–128, 1994.

[11] Young, S. et al. “HTK - Hidden Markov Model Toolkit”, Manual, 2006. Online: http://htk.eng.cam.ac.uk/

[12] Theodoris, S. and Koutroumbas, K., “Pattern recognition”, Academic Press, 1998.