A BOTTOM-UP APPROACH FOR HANDLING UNSEEN TRIPHONES IN LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Xavier Aubert, Peter Beyerlein, Meinhard Ullrich

Philips GmbH Research Laboratories Aachen, Weißhausstraße 2, D-52066 Aachen, Germany
E-mail: {aubert,beyerlei,ullrich}@pfa.research.philips.com

ABSTRACT

This paper presents an extension of bottom-up state-tying towards improved handling of unseen triphones. As opposed to the usual backing-off to diphones and monophones, the current method aims at finding a triphone model that has proven to exhibit some similarity with the unseen triphone. It is based on a probabilistic mapping of unseen contexts to clusters of triphone-states observed in the training data.

This algorithm has been applied to dictation tasks for three languages with vocabulary sizes ranging from 20k to 64k. The results compare favorably with those obtained using standard back-off rules. This technique also offers an alternative to top-down decision-tree procedures which are frequently used especially for their generalization capabilities.

1. INTRODUCTION

In large-vocabulary continuous speech recognition, context-dependent phone models are imperatively needed to cope with coarticulation effects. Triphone models are widely used as they capture the most important left and right influences. However, even when using large acoustic databases, many triphones that occur during recognition have not been observed in the training data [1]. The problem of these so-called "unseen" triphones is particularly acute when modeling cross-word contexts but appears also to be important for within-word triphones when decoding proceeds in very large vocabularies that include rare or proper names.

The key problem addressed in this paper is to supply a bottom-up state-tying technique with generalization capabilities, i.e. with the ability of finding a "good" model for an unseen context. The usual way of handling rare contexts with bottom-up state-tying consists in backing-off to less specific models like diphones or monophones [2]. This strategy leads to models with a lower accuracy, particularly when resorting to monophones that are obtained by pooling rarely seen contexts together. Moreover, it requires some careful tuning for selecting the most useful contexts based on a minimum number of occurrences in training script [3].

So far, this generalization problem has been mainly solved using decision-trees [4] [1] [5] that proceed "top-down" by going from less- to more specific contexts through successive (binary) splitting steps. In contrast, bottom-up techniques start from the most detailed level and aim at building robust entities by means of a merging process typically based on agglomerative clustering.

Decision trees are constructed data-drivenly guided by linguistic questions that introduce some a priori phonetic knowledge. This elegant and powerful technique has proven to yield impressive results [5]. However, the top-down split strategy might lead to a far from optimal classifier because of premature data fragmentation [1]. Furthermore, generalization follows from using binary 'yes/no' questions that always lead to a leave i.e. a particular model. Hence, the quality of the generalization partly depends on the nature of the questions. For example, if all questions about the immediate left and right contexts involve a single phoneme (not a group of similar phones), decision trees produce on a data-driven fashion nothing but generalization in terms of left- or right-diphones and monophones.

This paper presents an extension of the now standard bottom-up state-tying technique [6] towards improved handling of unseen triphones. The basic idea is to exploit the information encoded into the occurrences of pairs of left (right) contexts that are grouped in the same cluster of triphone states for predicting the cluster membership of a new context. A similar approach has been described in [7] for improving the coverage of trained triphone models on unlimited vocabulary test-sets. However, diphones and monophones are still needed to provide back-off models. Besides, no evaluation of this technique has been given. Our new algorithm is fully data-driven and no longer requires considering less specific contexts like diphones or monophones whatsoever. Any triphone context occurring in the training data is taken into account and all acoustic models are designed for groups of triphone states with specific left and right contexts.

The current paper is organized as follows. The algorithm producing clusters of tied states is first reviewed. Next, the extensions made in both training and decoding procedures

are explained. In particular, we present the construction of a probabilistic mapping of unseen contexts to clusters of triphone states being observed in the training data. Results are given for 3 languages with vocabulary sizes ranging from 20k to 64k and comparisons are made with our "best-tuned" results obtained with standard diphone and monophone backoff rules.

2. ALGORITHM DESCRIPTION

2.1. Review of State-Tying Technique

Referring to [6], state tying aims at producing homogeneous clusters of HMM states such that model complexity is balanced against the amount of data available for training. For each base phone, all corresponding triphone states are subjected to agglomerative clustering using (typically) a furthest neighbor criterion. Tying is thus restricted to triphone-states relevant to the same central phone and occupying the same position in the HMM topology. This process is controlled with two parameters, namely, a maximum cluster size and a minimum number of observations per tied state.

In the Philips system [2], state-tying is achieved in two consecutive steps. First, individual triphone states are modeled with a single Laplacian density and hierarchically clustered based on the following distance :

$$d(K_i, K_j) = \max_{m_k \in K_i} \max_{n_l \in K_j} \sum_c (|m_{k_c} - n_{l_c}|) \qquad (1)$$

where the clusters K_i and K_j are defined by a set of mean vectors m_k and n_l respectively, scaled with a globally pooled deviation vector. When the minimum value of $d(K_i, K_j)$ taken over all admissible (K_i, K_j) pairs, falls below a given threshold, both clusters are merged and their sets of means are joined together. Second, the total occupation count of each cluster of states is checked with respect to a minimum value and if smaller, that cluster is merged with its nearest neighbor according to distance (1).

After clustering being completed, a continuous mixture density HMM is estimated for each tied state by embedded Viterbi training [8].

2.2. Training of Rarely SEEN Triphones

This tying algorithm has been shown to significantly improve our acoustic modeling by allowing thrice as many triphones being selected and robustly trained [2]. However, rarely seen triphones (with typically 10 occurrences or less) appeared to be not properly handled as they led to non homogeneous clusters with degraded recognition performance. This is partly due to the poor estimation of their mean vector, inherent to the lack of data for such "sparse" acoustic events, but also due to the furthest neighbor criterion which seems inappropriate in this case. This problem has been greatly alleviated by considering a different metric for pairs of clusters involving one "sparse" cluster K_s (made of one or several rare triphone states only) and one non-sparse cluster K_j :

$$d(K_s, K_j) = \max_{m_r \in K_s} \min_{n_l \in K_j} \sum_c (|m_{r_c} - n_{l_c}|) \qquad (2)$$

According to (2), a rarely seen triphone state might be clustered based on the L_1 distance between its (single) mean vector m_r and the *nearest* mean vector characterizing a non-rare state in the second cluster K_j. Under these specified conditions, this metric does not lead to elongated clusters typical of the "chaining effect" that motivates the use of a furthest neighbor criterion (see [9], pp. 233). In our system, this modification leads to improved mixture density modeling of the obtained tied states as indicated by average likelihood score values.

2.3. Decoding with UNSEEN Triphones

Using the training procedure described above, a limited number of mixtures (see 3.3) succeeds in modeling the many thousands of seen triphones, a great deal of which occurs less than 5 or 10 times. However, unseen triphones have to be mapped in some way to appropriate acoustic models so that the system can successfully decode words carrying new phonetic contexts.

2.3.1 Probabilistic mapping of unseen triphone states

Let $c\{l+r\}s$ denote a triphone state where c is the central phone, l and r are resp. the left and right phonetic contexts and s is the state index. Our acoustic models are 3-states left-to-right HMMs, hence $s \in \{1, 2, 3\}$.
Let Tr be the set of all possible triphone states and Tr^o the set of triphone states *observed* in the training data. The set of *unseen* triphone states Tr^u is then given by : $Tr^u = Tr \backslash Tr^o$.
The outcome of the tying process is a partition of Tr^o in terms of N_K clusters K_i, $i = 1, N_K$ where each seen triphone state tr^o belongs to one and only one cluster $K_j \subset Tr^o$.

Our solution consists in searching among the available set of tied states for those that correspond best to the unseen triphones. More precisely, an unseen triphone state tr^u is assigned to one of the trained clusters $\hat{K}$ according to the probabilistic rule :

$$\hat{K} = \arg\max_K P(K \mid tr^u) \qquad (3)$$

and the corresponding tied-state HMM of $\hat{K}$ is used for tr^u. The probability of cluster membership $P(K|tr^u)$ can be expressed in terms of triphone elements :

$$\begin{aligned} P(K|tr^u) &= \sum_{tr \in Tr} P(K, tr|tr^u) \\ &= \sum_{tr^o \in Tr^o} P(K, tr^o|tr^u) + \sum_{tr' \in Tr^u} P(K, tr'|tr^u) \\ &= \sum_{tr^o \in Tr_o} P(K|tr^o, tr^u) \cdot P(tr^o|tr^u) + \hat{C}^u_{K,tr^u} \end{aligned}$$

where $\hat{C}^u$ represents the (unknown) contribution of all unseen triphones and is supposed to have a negligible influence on $\hat{K}$ as given by (3).

Assuming that the cluster membership is deterministic for each seen triphone tr^o, $P(K|tr^o, tr^u)$ is either zero or one and (3) becomes :

$$\begin{aligned} \hat{K} &= \arg\max_K \sum_{tr \in K} P(tr \mid tr^u) \\ &= \arg\max_K \sum_{tr \in K} P(tr\ ,\ tr^u)\ , \end{aligned} \tag{4}$$

the factor $1/P(tr^u)$ being removed as it is independent of K.

Given two triphone states $tr = c\{l+r\}s$ and $tr' = c\{l'+r'\}s$ related to the same center and state, $P(tr, tr')$ represents a similarity measure that can be further factorized into

$$\begin{aligned} P(tr\ ,\ tr') &= P_{c,s}(r, l, r', l') \\ &= P_{c,s}(r, r' \mid l, l') \cdot P_{c,s}(l, l') \\ &\approx P_{c,s}(r, r') \cdot P_{c,s}(l, l') \end{aligned} \tag{5}$$

This last approximation rests upon the assumption that left and right contextual influences are independent of each other which is of course not exactly right. However, this simplification is needed to achieve generalization capabilities and is somewhat analog to the use of questions concerning the left (right) only contexts for constructing decision trees [4]. The probabilities in (5) can be empirically estimated based on the counts of contexts co-occurring in the clusters. For example, using $*$ as the wild-card symbol standing for any context :

$$\hat{P}_{c,s}(r, r') = \frac{1}{\#Tr^o} \sum_K \left\{ \frac{1}{\#K} \Big[\sum_{tr \in K, tr = c\{*+r\}s} 1 \Big] \cdot \Big[\sum_{tr' \in K, tr' = c\{*+r'\}s} 1 \Big] \right\} \tag{6}$$

where $\#Tr^o$ is the total number of observed triphone states and $\#K$ the number of triphone states in cluster K.

2.3.2 Practical implementation & added heuristics

Several additional heuristics have been introduced for performing the first experiments described in the next section. Their exact impact is not known yet and is being systematically investigated.

- In (4), the sum has been replaced by the maximum.
- A positive bias has been introduced to emphasize the role of the left (right) context when handling the first (last) state of an unseen triphone.
- The probability values of cluster membership in (4) are subjected to a minimum threshold to ensure that decisions are based on a significant number of observations.
- When it is not possible to find a "host" cluster satisfying the latter constraint, the dependency upon the central phone in (5) is relaxed by taking a weighted average over all base phones : $P_s(r, r') = \sum_c P_{c,s}(r, r') \cdot P(c)$ and similarly for the state dependency.
- In case of failure, a set of broad phonetic rules is eventually applied to assign the unseen triphone to the least different observed context. Note however that in our experiments this ultimate choice has never been used as it was unnecessary.

Here follow typical examples of generalization automatically produced by our system when applied to American English. Among all unseen triphones arising while decoding, the algorithm succeeds in finding for 90% of the cases a cluster that includes either left or right phone context of the missing triphone. In this case, generalization thus consists in replacing only one, left *or* right, symbol with a context that has proven to be similar for the same base phone and state. Other non-trivial generalizations involving distinct left and right contexts are given in the table below together with word examples carrying the respective unseen and seen triphones.

Table 1: Examples of triphone-state generalizations

UNSEEN → SEEN	REC-Wrd	TRN-Wrd
y{d+un}1 → y{b+uh}1	in**d**ian	dub**i**ous
ee{p+aw}1 → ee{v+w}1	pe**o**ria	heav**y**weight
b{un+ah}2 →b{ih+aw}2	lauten**b**ach	anti**b**ody
t{p+oo}2 → t{k+oh}2	nep**t**une	fac**t**o
ow{m+th}3→ ow{l+d}3	m**ou**th	l**ou**d
oo{m+n}3 → oo{g+d}3	telem**u**ndo	g**oo**d

3. EXPERIMENTAL RESULTS

The present method has been applied to large-vocabulary continuous-speech recognition for 3 different languages, namely, US-English, French and German. The experiments have been carried out on dictation tasks defined within the SQALE project and the new results have been compared to those previously obtained [3]. No cross-word triphones have been considered so far.

3.1. Acoustic Training Databases

Table 2 below provides information about the phonetic richness of the databases used for training the acoustic models.

Table 2: Databases for acoustic training

LANGUAGE Database	US-Engl. WSJ	FRENCH BREF	GERMAN Phondat
#Speech Hours	12 H	8 H	12 H
#Spoken Words	131K W	90K W	125K W
#Differ.Words	9,084 W	13,850 W	1,725 W
#WW-Triphones	7,900 t	8,000 t	3,750 t

Although the total number of spoken words is comparable,

the number of distinct words and observed within-word triphones is much smaller for German. This follows from the content of "Phondat" made of a limited set of phonetically balanced sentences.

3.2. Test Conditions

The main characteristics for testing are summarized in table 3 below. To achieve similar Out-Of-Vocabulary (OOV) rates, it was necessary to consider a 64k German lexicon due to inflections and word compoundings. Test-set perplexities are computed with the trigram language models provided within SQALE. The number of within-word triphones contained in the recognition vocabularies (see last row of table 3) gives an idea of the number of unseen triphones, when compared to the corresponding figures in table 2. It ranges from 1,500 for French to more than 15,000 for German.

Table 3: Main characteristics of decoding tasks

LANGUAGE Newspaper	US-Engl. WSJ	FRENCH Le Monde	GERMAN Frankf. Rund.
Lex. Size	20K	20K	64K
OOV Rate	1.5%	1.7%	1.85%
TRIG Perp	132	116	271
#WW-Triph	11,700 t	9,400 t	19,250 t

3.3. Comparison of Triphone Modeling

The new training procedure involving all observed within-word triphones has been run for each language with the same maximum cluster size and same minimum number of observations, set to 100. The essential figures are presented in table 4 . When comparing the new acoustic modeling with the previous one, it appears that the number of tied states (and of densities) are quite similar though the number of states before tying is much larger. Looking at the word error rates, the new results appear slightly however consistently better. Keeping in mind that previous acoustic models had been tuned by trial and error on a development set, the new technique thus compares favorably.

4. CONCLUSION & FUTURE WORK

As a first advantage, the new technique makes training easier and more reliable by getting rid of any context selection rule based on minimum number of occurrences. Next, the algorithm is able to correctly handle all unseen triphones needed during decoding and achieves performances that are at least as good as our "best tuned" results obtained with the standard diphone/monophone back-off rules. Our current work aims at further improving the generalization capabilities and at comparing them with the (popular) decision tree technique, especially for handling cross-word triphones.

Table 4: Comparison of unseen triphone modeling

LANGUAGE Newspaper	US-Engl. WSJ	FRENCH Le Monde	GERMAN Frankf. Rund.
Diphone and Monophone Backing-Off [3]			
#Selec.States	7,315	4,408	3,870
#Tied States	3,157	2,456	2,858
Word Err-rate	14.7%	16.1%	19.7%
Data-driven Prediction of Unseen Triphones			
#Obser.States	23,763	24,462	11,032
#Tied States	3,326	2,603	2,785
Word Err-rate	14.4%	15.3%	19.2%
Rel.Improvemt	-2.2%	-4.6%	-2.7%

5. REFERENCES

1. Hwang M.-Y., Huang X., Alleva F.: "Predicting Unseen Triphones with Senones", Proc. ICASSP'93, II, pp. 311-314, Minneapolis, USA.
2. Dugast C, Beyerlein P, Haeb-Umbach R, "Application of Clustering Techniques to Mixture Density Modeling for Continuous Speech Recognition", Proc. ICASSP'95 pp. 524-527, Detroit, USA.
3. Dugast C., Aubert X., Kneser R., "The Philips Large-Vocabulary Recognition System for American-English, French and German", Proc. Eurospeech'95, pp. 197-200, Madrid, Spain, September 1995.
4. Bahl L.R., de Souza P.V., Gopalakrishnan P.S., Nahamoo D., Picheny M.A. : "Decision Trees for Phonological Rules in Continuous Speech", Proc. ICASSP'91, pp. 185-188.
5. Young SJ, Odell JJ, Woodland PC: "Tree-based Tying for High Accuracy Acoustic Modelling", Proc. SLT-HLT'95 Workshop, pp. 286-291, USA.
6. Young SJ, Woodland PC: "The Use of State Tying in Continuous Speech Recognition", Proc. Eurospeech'93, pp. 2203-2206, Berlin.
7. Digalakis V, Weintraub M, Sankar A, Franco H, Neumeyer L, and Murveit H, "Continuous Speech Dictation on ARPA's North American Business News Domain", in *Proceedings of the Spoken Language Systems Technology Workshop*, pp. 88-93, Austin, Texas, USA, 1995.
8. Steinbiss V., Ney H., Essen U., Tran B.-H., Aubert X., Dugast C., Kneser R., Meier H.-G., Oerder M., Haeb-Umbach R., Geller D., Hoellerbauer W., Bartosik H., "Continuous speech dictation - From theory to practice", in Speech Comm., Vol 17(1-2) pp. 1-20, August 95.
9. Duda R.O. and Hart P.E., "Pattern Classification and Scene Analysis", Wiley Interscience, New York, 1973.