USING AUTOMATICALLY-DERIVED ACOUSTIC SUB-WORD UNITS IN LARGE VOCABULARY SPEECH RECOGNITION

Michiel Bacchiani *Mari Ostendorf*

Electrical and Computer Engineering Department, Boston University, Boston, MA, USA

ABSTRACT

Although most parameters in a speech recognition system are estimated from data, the unit inventory and lexicon are generally hand crafted and therefore unlikely to be optimal. This paper describes a joint solution to the problems of learning a unit inventory and corresponding lexicon from data. The methodology, which requires multiple training tokens per word, is then extended to handle infrequently observed words using a hybrid system that combines automatically-derived units with phone-based units. The hybrid system outperforms a phone-based system in first-pass decoding experiments on a large vocabulary conversational speech recognition task.

1. INTRODUCTION

Large vocabulary speech recognition systems typically represent lexical entries in terms of sub-word units, for which acoustic models can be reliably estimated. Part of the system design is therefore to decide on a suitable unit inventory and define the mappings from lexical entries in the vocabulary to linear strings or networks of units (i.e. define the lexicon). Although the parameters of the unit models are generally estimated from data using an objective function such as maximum likelihood, no such function is used in the unit inventory and lexicon design – the problem is typically simplified by using phone-based units and a hand-crafted lexicon. As a result, the unit inventory and lexicon are unlikely to be optimal in terms of the objective function used throughout the design of the rest of the system. Furthermore, the standard dictionaries available usually characterize citation-form pronunciations, which are unlikely to be reflective of the realization of many words in spontaneous conversational speech.

An alternative to manual derivation of a unit inventory and lexicon is to learn them from data. We will refer to a unit derived in this way as an Acoustic Sub-Word Unit (ASWU). A number of researchers have investigated this approach, e.g. [1, 2, 3, 4], but decoupling the unit inventory and lexicon design problems, which are clearly related. This limitation can be addressed by iterative reestimation of the acoustic model and the pronunciations, as in [5]. Another problem is that for speaker-independent tasks, where there is a lot of variation in the initial acoustically-motivated labeling of tokens, it is expensive to determine the optimal pronunciation among the large number of candidates and difficult to rule out cases when the vast majority occur only once. Furthermore, the optimal pronunciation may not even be among the observed candidates if the unit inventory is large, as would be the case for large vocabulary recognition. Not surprisingly, most results are reported on small tasks and/or compared to context-independent phonetic models. To address the large inventory problem, we proposed a method of joint lexicon and unit design that imposes pronunciation consistency constraints from the early stages of the design process to achieve a match between the unit inventory and the lexicon [6]. Experimental results on the Resource Management task show performance comparable to or better than phone-based systems using state-of-the-art clustered triphone models [7]. However, the approach assumes that several training tokens are observed for each word in the lexicon. Thus, as proposed, the algorithm does not address the needs of most large vocabulary speech recognition tasks, where many words in the lexicon are observed infrequently (or not at all) in training.

This paper describes an extension of previous work that uses a hybrid system of automatically-derived units together with phone-based units (clustered triphone states) for large vocabulary, conversational speech recognition. The approach combines the advantages of automatically derived units and pronunciations for frequently observed words with the generalizability of phonetic units for infrequent words. While the longer term goal might be to develop a technique for automatically deriving units that are general, it is likely that much of the possible gain is achievable by a hybrid system. In the Switchboard corpus, for example, the top 400 words cover 87-88% of the training corpus in terms of word tokens, and each is observed more than 100 times.

The remainder of the paper is organized as follows. The hybrid system design methodology is described in section 2, including details about the sub-system design parameters and issues associated with merging them. Speech recognition experiments on the Switchboard corpus are provided in section 3. This paper addresses the problem of generalization; the remaining problem of pronunciation variability for automatic unit design on conversational speech is discussed in section 4.

2. SYSTEM DESIGN

The training of the hybrid system involves first separately designing a phone-based system on the full corpus and automatically-derived units on the subset of the corpus covered by the most frequent words. All units in both cases are represented using a hidden Markov model (HMM). The training for the automatic unit part of the system is essentially the same as that described in [6], but is reviewed in section 2.1 for completeness. Training for the phonetic unit sub-system, described in section 2.2, uses standard data-driven HMM triphone clustering techniques, with the exception that state tying is unconstrained (as in the ASWU case). The two sub-systems are combined in parallel for the most frequent words, then pronunciations are pruned and reestimated to form the hybrid system, as discussed in section 2.3.

2.1. ASWU Sub-System Design

The two basic algorithmic steps of all the proposed unit inventory design algorithms are an acoustic segmentation followed by a clustering step to define the unit inventory. The key elements that differ in our approach are the use of pronunciation-related constraints in both steps of the design algorithm, as well as the consistent use of a maximum likelihood objective function.

The first step in designing an ASWU system is **acoustic segmentation**, that is, finding segmentation times that divide each word token into piecewise stationary regions that can be reasonably well modeled with a single HMM state.[1] Unconstrained acoustic segmentation [1] involves recursive updating for every time t and every allowable number of segments n:

$$\delta(t,n) = \max_{\tau} \left[\delta(\tau-1, n-1) + \log \mathrm{p}(x_\tau, \ldots, x_t | \mu_{\tau,t}, \Sigma) \right],$$

where τ may have some duration constraints and $\mathrm{p}(\cdot|\mu_{\tau,t}, \Sigma)$ is a generalized likelihood computed using a multivariate Gaussian model with a known diagonal covariance Σ (the grand variance of the entire training corpus). In our implementation, we start with fixed word begin and end times and introduce the constraint of a fixed number of segments within each word token, where the number is equal to the median length for that word using an unconstrained segmentation. (The word boundary times are provided by a phone-based HMM system.) Using a fixed number of segments per word is equivalent to the linear pronunciation model used for the vast majority of words in most speech recognition systems.

The second step involves **clustering** the results of the segmentation step to define the unit inventory. Again a pronunciation consistency constraint is introduced. Before clustering, the data is grouped, computing the sufficient statistics for each collection of segments originating from different training tokens in the same position within a lexical entry. The sufficient statistics for the Gaussian mean model are the sample mean and covariance and the total number of vector observations contained within the group. These sufficient statistics are stored for each unique position within each unique lexical entry: if there are V entries in the vocabulary and the median pronunciation length is R, the data is grouped into VR groups. Clustering involves a combination of divisive and K-means clustering of these groups, almost as if they were individual data points. The sufficient statistic representations of these atomic groups cannot be split in clustering, thus ensuring the pronunciation consistency.

The clustering algorithm used here also differs from that used in [1, 5] in that maximum likelihood is used as an objective rather than minimum Euclidean distance. Specifically, the repartitioning step involves computing the likelihood of segments given the model parameters of a cluster, i.e. a negative log likelihood "distance". The cluster reestimation procedure consists of finding the maximum likelihood parameter estimates of a Gaussian distribution from the data contained in the cluster. Cluster centroids therefore directly represent unit models and clustering addresses both the inventory and model design problems, whereas in other work unit model parameters had to be estimated in a separate step from the data partition defined by clustering.

The models derived by clustering are then reestimated to use mixture distributions, using the incremental mixture-splitting technique described in [8].

2.2. Phone-Based Sub-System Design

A hand-crafted phonetic unit inventory (54 phones) and lexicon are used for the phonetic part of the system. The phonetic unit models are left-to-right 3-state HMMs with a topology allowing the center state to be skipped. The transition probabilities of these HMMs were kept uniform. The state emission probabilities were modeled by mixture Gaussian distributions, with distributions tied using triphone clustering. Parameters for these models were estimated from data by gradually increasing the system complexity, starting from single Gaussian distribution models for context-independent phonetic units as described below.

First, initial context-independent phone model parameters were estimated from a phone-level segmentation provided by another speech recognition system. The parameters of these models were then refined using Viterbi training with the constraint of fixed word boundaries. Next, the model inventory is increased by explicit modeling of context at the state level, where the resulting models are referred to as tri-states. Single Gaussian emission probabilities are estimated for all unique tri-state units by Baum-Welch reestimation. The system complexity is then further increased by modeling context at the phone level. The tri-state system is used to realign the data, and sufficient statistics are computed for states in all unique left and right phone contexts but ignoring contexts across word boundaries. These sufficient statistics were then clustered using combined divisive and K-means clustering with likelihood as the objective function, similar to the ASWU sys-

[1] In fact, the algorithm is not restricted to HMMs and has also been implemented for polynomial mean trajectory segment models [3].

tem. No structure was imposed on clustering, so units are allowed to share across center phone identity and state position within the phone models. The cluster inventory was initialized with the tri-state model inventory derived previously by the Baum-Welch reestimation. The shared triphone state distributions are then refined by 3 iterations of Baum-Welch reestimation. The complexity of the clustered triphone models is then increased by estimating mixture distributions, using the incremental mixture-splitting technique described in [8].

2.3. Hybrid System Design

The hybrid system integrates the two types of units. One option for system building is to simply join the two types of unit systems trained independently. In such a system, the unit inventory is defined as the union of the automatic unit and phonetic unit inventories, and the lexicon uses automatic units for the most frequent words and phonetic units for the remaining entries. Although this approach to the hybrid system design is simple, it has several disadvantages. First, the automatic units might provide more accurate models in comparison to the phonetic unit word models for some of the most frequent words but possibly not all of them. To solve this problem, a criterion is needed to decide whether to include the phonetic or automatic unit pronunciation in the lexicon. Second, as the most frequent words are now modeled by the automatic units, the phone-based units do not need to cover the full space of triphones and the parameters can be reestimated. However, eliminating a large portion of the training data could make the models less general and therefore less accurate on unseen data. Third, as the automatic units were trained using isolated tokens with fixed word boundaries, it is likely that embedded training (allowing the word boundaries to shift) will result in more accurate model parameter estimates.

As an alternative to simply joining the independently-trained automatic and phonetic unit systems, a parameter reestimation step is performed on a parallel version of the hybrid system. In other words, the most frequent words are represented with multiple pronunciations: the automatic unit pronunciation and the phonetic unit pronunciation. The estimation step of the Baum-Welch algorithm is then used to compute two probabilities: the standard "state occupancy" counts (the probability of being in a particular state at a particular time given the whole of the observation sequence), and the "word-initial state transition counts" (the probability of transitioning into either the first automatic unit or phonetic unit states of a multiple pronunciation word at any time given the whole observation sequence). The word-initial state transition counts are used to estimate the probability of each possible pronunciation, which can be used to weight or to prune the different pronunciation alternatives for each word. The state occupancy probabilities can be used to reestimate the model parameters of either or both types of unit models. Here, only the ASWU models are estimated to avoid possible problems associated with triphone model reestimation from a biased data sample for unseen models. In addition, our implementation uses pruning rather than weighting, in which case pronunciations are removed from the lexicon. Therefore, it is useful to do a second estimation pass with the single pronunciations.

3. EXPERIMENTS

Experiments were performed on the Switchboard conversational speech corpus. The training set consisted of approximately 120 hours of speech from about 2500 conversations. Feature vectors consisting of 14 gender-dependent vocal-tract-length (VTL) normalized cepstral coefficients and derivatives were available were computed at a rate of 100 vectors per second. The VTL-normalization was performed using a segment level equivalent of the frame level technique described in [9]. A test set of approximately 30 minutes of speech from 7 conversations was defined. Gender detection was performed using the likelihoods of the VTL normalization model. The test lexicon contained 20383 entries, and a bigram language model was used in decoding. The OOV rate for the test set using this lexicon was 0.7%.

The **phone-based system** was trained using the method described in Section 2.2, resulting in clustered triphone inventory sizes of 5265 and 5244 for the male and female system, respectively. The triphone state emission probability distributions were refined further to 12-mixture distributions by mixture splitting and Baum-Welch training. The 12 mixture model inventory was then used to resegment the training data allowing both phone state as well as word boundaries to move with up to 400 frames. In addition, at each word boundary, an optional silence word was allowed to be inserted. The obtained word level segmentation contained explicit information on which pronunciation variant was used for those lexicon entries having multiple pronunciations.

The **ASWU sub-system** was trained as described in section 2.1, with an initial acoustic segmentation tuned so that there were on average 3.4 acoustic segments per phone segment. A total of 3000 automatic unit models/gender were estimated through constrained clustering. The number was chosen arbitrarily to be a large fraction of the phone-based system number, since the models covered a small percentage of the tokens but a large percentage of the data that the phone-based models covered. Again, 12 mixture distributions were estimated for the automatic unit models by mixture splitting and Baum-Welch training. The parameters were estimated from the word tokens of the 400 most frequent words, keeping word boundaries fixed to the times provided by the phone-based system.

For the **hybrid system**, one Baum-Welch reestimation pass was performed. A histogram of the pronunciation probabilities for the automatic unit pronunciations of the most frequent lexicon entries as estimated by the reestimation pass is given in figure 1, showing that the automatic unit pronunciation is more likely (i.e. a better fit to the data) for most but not all words. For cases where there are multiple phone-based pronunciations per word, there are the same number of automatic unit pronunciations and these probabilities are summed in the figure.

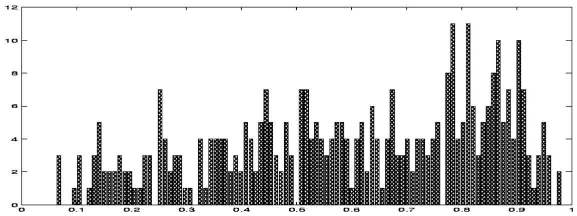

Figure 1: Histogram of estimated probabilities of the automatic unit pronunciations of the 368 most frequent words (female).

Results are illustrated for the female models; similar behavior is observed for the male models. A pruned lexicon was then obtained by using the automatic unit pronunciations when they had combined probability larger than 0.5, and the phone-based pronunciations otherwise. The automatic unit model parameters were then reestimated by another Baum-Welch reestimation pass, but the phonetic model parameters were held fixed.

Recognition performance on the test set using the 12 mixture phone-based models was 40.0% accuracy. Using the hybrid system with the pruned lexicon without reestimating the unit model parameters gave a 40.5% accuracy. After a single pass of Baum-Welch training, the pruned hybrid models gave a 41.9% accuracy.

4. DISCUSSION

In summary, this paper describes a hybrid system for large vocabulary speech recognition that combines the advantages of automatically-derived acoustic units for high frequency words with the advantages of generalizable phone-based units for infrequently observed words. Experimental results on the Switchboard corpus show that the automatically-derived units and associated pronunciations do indeed give a better fit to the data than phone-based units, in terms of higher training likelihood and improved recognition accuracy. The experimental results are on a first-pass decoding paradigm (word-internal triphones and bigram language model), and further system development and experiments are needed to demonstrate improvement in a more complex, multi-pass decoding system. In our experiments, the phone-based models are based on the full data set, and are not retrained to reflect the fact that they are not used for the most frequent words. One unresolved question is whether there is a performance gain to be had from retraining or adapting the phone-based models.

For applications involving recognition of spontaneous speech, a limitation of the algorithm described here is the assumption of a single linear pronunciation per lexical item. In spontaneous speech, phone segments can be modified dramatically and are frequently dropped completely [10], so that a single linear pronunciation is likely to be inadequate for representing many words. The problem is addressed here by pre-specifying multiple ASWU pronunciations according to whether a word in the lexicon had multiple phonetic pronunciations. However, one would ideally like to learn the pronunciation variants automatically from data as well, as explored in other ASWU work [11]. A straightforward extension of the temporal sequential unit splitting algorithm to parallel unit splitting would allow for multiple pronunciations within the context of unit design with pronunciation constraints, assuming a redefinition of atomic units. A related problem is the modeling of cross-word phonological affects. which recent work has addressed by defining multi-word lexical entries with different pronunciations. Given an algorithm for learning multiple ASWU pronunciations, it is straightforward to combine this with multi-word lexical entries and can be combined with an acoustically motivated definition of the multi-word set [3].

Acknowledgments: This work was supported by ATR Interpreting Telecommunications Laboratory.

5. REFERENCES

1. K.K. Paliwal, "Lexicon building methods for an acoustic sub-word based speech recognizer," *Proc. Int. Conf. on Acoust., Speech and Signal Proc.*, vol. 2, pp. 729-732, 1990.
2. L. R. Bahl, P.F. Brown, P. V. de Souza, R. L. Mercer and M. A. Picheny, "A Method for the Construction of Acoustic Markov Models for Words," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 4, pp. 443-452, 1993.
3. M. Bacchiani, M. Ostendorf, Y. Sagisaka and K. Paliwal, "Design of a Speech Recognition System based on Non-Uniform Segmental Units," *Proc. Int. Conf. on Acoust., Speech and Signal Proc.*, vol. 1, pp. 443-446, 1996.
4. T. Svendsen, F. K. Soong and H. Purnhagen, "Optimizing baseforms for HMM-based speech recognition," *Proc. European Conf. on Speech Commun. and Technology*, vol. 1, pp. 783-786, 1995.
5. T. Holter and T. Svendsen, "Combined Optimisation of Baseforms and Model Parameters in Speech Recognition Based on Acoustic Subword Units," *Proc. IEEE Workshop on Automatic Speech Recognition*, pp. 199-206, 1997.
6. M. Bacchiani and M. Ostendorf, "Joint Acoustic Unit Design and Lexicon Generation," *Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition,* 1998, pp. 7-12.
7. P. C. Woodland and S. J. Young, "The HTK tied-state continuous speech recogniser," *Proc. European Conf. on Speech Commun. and Technology*, vol. 3, 1993, pp. 2207-2210.
8. S. J. Young and P. C. Woodland, "The Use of State Tying in Continuous Speech Rcognition," *Proc. European Conf. on Speech Commun. and Technology*, vol. 3, 1993, pp. 2203-2206.
9. S. Wegmann, D McAllaster, J. Orloff and B. Peskin, "Speaker Normalization on Conversational Telephone Speech", *Proc. of the Int. Conf. on Acoust., Speech and Signal Proc.,* , pp. 339-341, 1996
10. S. Greenberg, "Speaking in shorthand – a syllable-centric perspective for understanding pronunciation variation," *Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition,* pp. 47-56, 1998.
11. T. Holter and T. Svendsen, "Maximum likelihood modeling of pronunciation variation," *Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition,* pp. 63-66, 1998.