A STUDY ON TASK-INDEPENDENT SUBWORD SELECTION AND MODELING FOR SPEECH RECOGNITION

Chin-Hui Lee, Biing-Hwang Juang, Wu Chou and Jose Javier Molina-Perez †

Multimedia Communications Research Lab.
Bell Laboratories, Lucent Technologies
600 Mountain Avenue
Murray Hill, NJ 07974, USA

ABSTRACT

We study two key issues in *task-independent* training, namely selection of a *universal* set of subword units and modeling of the selected units. Since no *a priori* knowledge about the application vocabulary and syntax was used in the collection of the training corpus and the recognition task is frequently changing, the conventional strategy can no longer provide the best performance across many different tasks. We present a new approach that use the complete sets of right and left context-dependent units as the basis phone sets. Training of these models is accomplished by a new training criterion that *maximizes phone separation* between competing models. The proposed phone selection and modeling approach was evaluated across different tasks in American English. Good recognition results were obtained for both context-independent and context-dependent phone models even for unseen tasks. The same strategy has also been applied to two other languages, Mandarin Chinese and Spanish, with similar success.

1. INTRODUCTION

It is well-known that hidden Markov model (HMM) based speech recognizers often degrade in performance when the testing conditions deviate from those of the training data (e.g. [5]). In this paper, we examine one of the least studied *mismatch* conditions, namely the training and testing mismatch due to vocabulary and task specification. *Task-dependent* (TDEP) training has been the prevailing strategy in speech recognition literature where the task specifications in testing is mostly identical to those in training.

For applications in which the target vocabulary is either not specified *a priori* or modified frequently from one task to another, a training procedure aiming at *task-independent* (TIND) phone modeling becomes necessary. Research in TIND training was pioneered in [2]. The goal of TIND training is to create a set of subword models that is capable of handling new tasks without the need of collecting new training materials, capable of generating a context rich set of subword units and models to handle new vocabularies and capable of producing a reasonable performance even for unseen tasks. Therefore training material design, phone set selection and phone modeling are three key research issues need to be addressed.

For training material design, it is customary to have a set of sentences that are phonetically rich and cover a large number of useful triphone context. In this study we investigate the subjects of unit selection and modeling. For phone set selection in TDEP training, the criterion is often based on frequency of occurrences, i.e. selecting the phone units that appear enough times in the training set (e.g. [3]). In addition to using the simple set of context-independent (CI) phone units in TIND training, one can also use the same abovementioned strategy in selecting context-dependent (CD) units to improve context coverage. However, this often results in poor performance due to the inefficient usage of the training data, i.e. some models that used up a large portion of data to train ended up not being used in a new task because their corresponding contexts do not appear in the vocabulary of the task (e.g. [4]). Another strategy to enhance the task-specific performance is the notion of *vocabulary learning* [2]. Instead of fixing a set of CD phone models at training time, one can use the context information of the target vocabulary and task grammar and select a new set of phone models to train for each new task. This was shown to produce a good performance by incorporating such task-specific context information [2]. In all TIND approaches maximum likelihood (ML) training is still the dominanting strategy.

Assuming the TIND training corpus is given, we present a new approach that use the *complete* set of *right* or *left* context-dependent units as the choice for a task-independent phone set. We also propose an adaptive *model composition* strategy to form a CD model for a triphone unit based on merging the first state of the corresponding left CD model and the last state of the corresponding right CD model into the first and the last state of the desired, but unseen triphone model. This allows creation of such models for new vocabulary words in a new task environment at *runtime*. We also present a new discriminative training to estimate phone models based on the criterion of minimum phone recognition error rate, as opposed to word recognition error rate, over the training data. Conventional minimum classification error (MCE) training techniques, which attempt at minimizing word or string recognition errors, require that the vocabulary in the task be defined *a priori* so that the recognition performance in terms of the word error rate can be properly evaluated. The proposed training algorithm attempts to achieve a maximum separation between phone models so that they are more robust for new tasks than the traditional modeling results which are opti-

† J. J. Molina-Perez is now with AT&T Laboratories.

mal only for the vocabulary specified at the training stage. The proposed approach is also highly efficient in that phone recognition is used as the first step as opposed to the conventional approaches in which word recognition is applied first in preparation for discriminative training. Since the databases used for TIND training often cover a long list of distinct words and a large number of utterances in various linguistic contexts, speech recognition by word modeling is extremely slow if not impossible to accomplish.

The proposed phone set selection strategy, adaptive triphone model composition technique and the MCE training method were evaluated on TIND training of American English, Mandarin, and Spanish phone units. We observed the following: (a) phone recognition rates were significantly improved for all three languages using either CI, right CD, or left CD units trained discriminatively; (b) the model prediction technique based on vocabulary learning of new phone units gave better performance than that obtained without using model prediction; and (c) the right CD set worked slightly better than the left CD set; both the right and left CD sets gave much better results than the CI set.

2. BASELINE SYSTEM AND DATABASES

The baseline system we used for training and recognition is described in detail in [3]. Input speech, sampled at 8 kHz, was initially pre-emphasized ($1-0.95z^{-1}$) and grouped into frames of 240 samples with a shift of 80 samples. For each frame, a Hamming window was applied followed by a 10^{th} order LPC analysis. A liftered 12-dimensional LPC-derived cepstral vector was then computed. The first and second time derivatives of the cepstrum were also computed. Besides the cepstral-based features, the log-scaled energy, normalized by the peak, and its first and second order time derivatives were also computed. Thus, each speech frame was represented by a vector of 39 features [3].

Except for the background silence unit, each subword unit is modeled by a 3-state left-to-right HMM with no state skip. Each state is characterized by a mixture Gaussian state observation density. Training is done with an iterative segmental ML algorithm (e.g. [4]) in which all utterances are first segmented into subword units. The Baum-Welch algorithm is then used to estimate the parameters of the mixture Gaussian densities [4] for all states of subword HMMs. Recognition is accomplished by a frame synchronous beam search algorithm [3] to determine the sequence of words (or phones) that maximizes the likelihood of the given utterance. The databases used in this study is now described briefly in the following.

2.1. Task-Independent Training Data

The database used for task-independent training is a set of 12,000 utterances of general phrases of American English collected by AT&T[1] over long distance public subscribers telephone network (PSTN). More than 2,000 talkers each speaking up to 7 phrases were included. These calls were originated from 100 different area-codes (also known as Number Plan Areas or NPAs) with 300-400 calls from each of the 7 major dialect regions of US. Female and male speakers were fairly represented (44% and 56% respectively). The database was recorded over a period of 3 months. This training set is referred to as GP93 in the remainder of this paper. The text material was selected from the AP Newswire. Each phrase is semantically correct with length ranging from 2 to 4 words. The selection of the phrases was based on a greedy algorithm such that a maximum triphone coverage is obtained.[2]. Over 6,000 distinct words were included in the recording.

2.2. Task-Dependent Testing Databases

For testing purposes, we have tried a number of tasks, including isolated word recognition of large sets of proper names. We have also tested a number of continuous speech recognition tasks of natural number recognition, date and time recognition, etc. In this paper, we studied two testing sets in detail: the first is 232 utterances of six alpha-digit strings (referred to as AD6) spoken in continuous speech mode by about 100 speakers, and the second is a subset of 2400 utterances of New Jersey town names (referred to as NJT), spoken in isolated phrase mode using either analog or ISDN telephones by about 100 speakers. A subset of 1219 town names is chosen as the NJT test vocabulary.

3. TASK-INDEPENDENT PHONE SETS

One critical issue in TIND subword modeling is the selection of a *universal* phone set that is easy to train and gives a good performance even for unseen tasks. Although the set of CI units is an obvious choice, CD units often give a better recognition accuracy.

3.1. CI Phone Set

The simplest way to obtain a set of task independent phone models is to choose the set of CI units, each of them modeling respectively the phoneme of the language. There is no context mismatch problem here. For American English, we used a set of 40 phonemes commonly used in the Bell Labs' Text-to-Speech grapheme-to-phoneme transcription rules.

3.2. Double-Context Phone Sets

Conventional selection (e.g. [3]) of a CD phone set based on the occurrence of the triphone context in the TIND training corpus is inappropriate because the target vocabulary may have a very different context coverage from that in the training data. If the context coverage of the testing data is very different from that of the training data, it is likely that only a small portion of the phone units are actually used in recognition for the particular task. This is undesirable because only a small portion of the limited training data is used effectively and it often results in a poor recognition performance (e.g. [4]).

3.3. Single-Context Phone Sets

In order to broaden the context coverage to deal with all unknown tasks and maintain the performance advantage of the CD units, we propose the use of the *complete* set of right or left CD phone units as the choice to form a universal TIND phone set. This also has the advantage that

[1] The Database was designed and recorded by Rich Sachs of the AT&T Voice and Audio Processing Architecture Department in Holmdel in 1993.

[2] The algorithm was graciously provided by Jan van Santen of the Linguistics Research Department of Bell Labs in Murray Hill

the set of models, once trained, remains fixed for all future tasks. No re-training is needed. Since not all single-context CD phone units appear in the training set and not all units appear frequently enough, we use a threshold of 50 to limit the number of single-context units. This resulted in 1207 right and 1260 left CD phone units (as opposed to the full set of 1640 units each). To deal with some unseen context, we also supplement the right and left CD phone sets with the set of CI units which brings the two sets to 1247 and 1300 units respectively. The models for the two sets are trained separately using the maximum likelihood training criterion. These two model sets can also be combined with other double-context model sets to form larger sets that cover a wider range of context. In our preliminary experiments, we found that the right CD phone set always slightly outperformed the left CD phone set although more units are included in the later set. We suspect that it is due to the reason that we perform speech recognition in a left-to-right manner and therefore the right CD phone units carry slightly more information than that of the left CD phone units.

4. ADAPTIVE TRIPHONE COMPOSITION

It is well known that the performance of a recognition task often depends on how well the set of CD units covers the context of the vocabulary and the grammar of the task. However for applications that the task definition is flexible and changing, it is not possible to retrain the set of CD models, every time a new task is encountered. One way to improve the context coverage is to start with a *fixed* set of CD units and add the task-dependent CD units and their corresponding models at runtime. We propose starting with the right (or left) CD set which is already context rich. We then add the triphone units that are required by the new task by examining the context coverage implied by the vocabulary words and the task grammars (if cross-word triphones are to be included). We rank the importance of each triphone by the number of occurrences in the vocabulary words and select the top Q of them to be included for the CD phone set of the particular task. This allows an adaptive and effective creation of unseen triphone models for modeling new context in a new task environment.

To form a CD model for each triphone unit, we propose a *model composition* algorithm by tying the first and the last state of the desired triphone unit with the first state of the corresponding left CD model and the last state of the corresponding right CD model respectively. The middle state of the new triphone can be derived from the middle state of either the left CD model, the right CD model, or the CI model with the corresponding context. We obtained slightly better performance using the right CD model. This is consistent with the fact the right CD set always outperformed the left CD set in all of our tests. Such a vocabulary learning procedure and model composition strategy are quite efficient because the new models are introduced through a *tied-state* structure which only slightly increases the computation load of the recognition system. It is also practical that no re-training is required. Only the context information about the vocabulary words needs to be provided for each new task.

5. MCE-BASED PHONE MODEL TRAINING

The conventional MCE training [1] aiming at minimizing word recognition error in the task-dependent training cases, no longer provides the best model set because the target vocabulary is not fixed in testing. To alleviate this difficulty, we propose replacing the minimum word error objective with a new minimum phone error objective. This allows us to use the same algorithm [1]. All we need to do is to first change the transcription of each utterance (by simply replacing each word with its corresponding phone transcription through the use of the training lexicon). We then perform phone recognition (instead of word recognition) when generating the N-best competing phone strings for each given utterance phone transcription ($N = 4$ in all of our experiments). For a typical phone set of less than 2,000 units, this is more efficient than word recognition because in a typical TIND training set, there is a long list of distinct words (usually more than 5,000) which makes it a very slow process to obtain the N-best competing word strings for each utterance during MCE training.

The model parameters are estimated using the segmental *generalized probabilistic descent* (GPD) algorithm [1]. The same algorithm can be used for both CI and CD model training. Usually we perform nine iterations of GPD training for CI models and 15 iterations of GPD training for CD models. Different size of CI and CD models have been created. For CI models, we have a maximum of 8, 16 or 32 mixture components for each HMM state. For the single- and double-context CD models, we have a maximum of 4, 8 or 16 mixture components per state.

6. EXPERIMENTAL RESULTS

We now reported on some experimental results to show the validity and potentials of our proposed approaches. First we show, in Table 1, phone and word error rates on training and testing data using four models sets with either 8 or 32 mixture components per state and trained using either the standard ML approach or the proposed MCE approach. It is clear that MCE models gave a much better performance than that obtained with ML models, especially for the training subset which contains 538 utterances of phrases. It is also clear that the 32-mixture model sets, significantly outperform the 8-mixture model sets. The best performance we have for the AD6 test data was 6.2% word error using MCE models trained on large corpora of (TDEP) connected digit and connected alphabet strings. It shows that TDEP training is still preferred to TIND training for small vocabulary (e.g. alphabets, digits, natural numbers, etc.) recognition in which the vocabulary words are somewhat fixed for different applications, and a large amount of training data can be collected and shared among many applications.

-	ML-8	MCE-8	ML-32	MCE-32
GP93	58.8	36.3	52.2	34.8
AD6	15.1	14.5	11.9	10.7
NJT	19.8	13.3	11.2	9.9

Table 1. Word error rates (%) using CI model sets.

Next we report on results using double-context CD units in Table 2. We list the number of CD phones, N, as a function of the count threshold, T (from 30 to 200), for the GP93 set. It can be seen that the more the number of CD units the less the amount of training data is used to model each unit. A maximum of 4 mixture components is used to model each state for all four cases. The word error rates using these CD model sets are also shown in Table 2. It indicates that the CD models are not as robust as the CI model sets when testing on data that contain very different context information (shown in the AD6 and the NJT rows). It also shows that the recognition performance often degrades when too many units are not well modeled. We can improve the robustness by including the CI models into the CD model sets (recognition results shown in the bottom row of Table 2).

T	200	150	50	30
N-CD	576	1132	2109	3062
AD6	16.3	13.9	15.2	20.5
NJT	13.6	13.1	10.3	17.2
+CI	11.2	10.7	10.4	11.5

Table 2. Word error rates (%) using double-context CD sets.

We now report on testing results using only single-context CD model sets in Table 3. For both AD6 and NJT testing, both the left and the right CD sets outperform the results shown in Tables 1 and 2. It also shows that the right CD set is better than the left CD set. When combining the left and right CD sets, there is a slight improvement for both testing sets (shown in the rightmost column of Table 3). Part of the reasons that single-context CD models are more robust than the double-context CD models is because the training data is more evenly distributed across different double contexts when they are merged into a single context. This is shown in Table 4 (M is the actual number of units used) that the percentage of HMMs used in NJT testing varies a lot from 99% using 576 CD units to 59% using 3,062 CD units while it remains at about 75-80% for most of the tasks we have tested using the right CD model set.

CD1 Train	Right (MLE)	Right (MCE)	Left (MLE)	Both (MLE)
AD6	13.4	12.2	13.7	13.2
NJT	6.4	5.8	6.8	6.2

Table 3. Word error rates (%) using single-context CD sets.

CD	576	1132	2109	3062	Right
M	568	978	1455	1795	804

Table 4. Number of HMMs used in NJT testing.

Finally the model composition strategy is tested for both single-context and double-context CD sets. In Table 5, we compare results with and without using additional triphone models obtained by model composition. Shown on the left is the CD-576 case. We choose to add the top 500 most frequently used triphones in the NJT vocabulary. Only 444 of them are actually added because the rest of the 56 overlaps with the original set. It indicates that the top 576 triphones in the training data and the top 500 triphones required in testing overlap only by about 10%. By adding the 444 models using the abovementioned model composition algorithm, a 23.5% error reduction was observed. Shown on the right of Table 5 are the results after adding 1520 top triphones to the right CD set. We only achieved a slight improvement of 11% error rate reduction. This agrees with our conjecture that the right (or the left) CD set has already a good starting context coverage that even adding a large number of task-specific triphone models does not improve the performance significantly.

CD	576	+444	right	+1520
NJT	13.6	10.4	6.4	5.7

Table 5. Word error rates (%) w/o model composition.

7. SUMMARY

We have discussed the two key issues of phone set selection and phone modeling in task-independent training. We presented a new strategy to use the set of single-context (left or right) phones as the *universal* phone set. This expands the context coverage yet maintains the modeling efficiency because most of the training data are used effectively in new task environments. We also proposed a new adaptive model composition strategy to generate useful triphone models from the existing sets of single-context and context-independent phone models for triphones not seen in training data. To improve modeling accuracy, we have used the MCE criterion in training which resulted in a better model separation and a better recognition performance. We have also applied the same strategy to recognition of Mandarin Chinese and Spanish with a similar degree of success.

REFERENCES

[1] W. Chou, C.-H, Lee and B.-H. Juang, "Minimum Error Rate Training Based on the N-Best String Models," *Proc. ICASSP-93*, pp. II-652-655, Minneapolis, 1993.

[2] H.-W. Hon and K.-F. Lee, "Vocabulary Learning and Environmental Normalization in Vocabulary-Independent Speech Recognition", *Proc. ICASSP-92*, pp. I-485-488, San Francisco, 1992.

[3] C.-H. Lee, E. Giachin, L. R. Rabiner, R. Pieraccini and A. E. Rosenberg, "Improved acoustic modeling for large vocabulary continuous speech recognition," *Computer Speech and Language*, Vol. 6, pp. 103-127, 1992.

[4] C.-H. Lee, J.-L. Gauvain, R. Pieraccini and L. R. Rabiner, "Large Vocabulary Speech Recognition Using Subword Units," *Speech Communication*, pp. 263-280, Vol. 13, Nos. 3-4, 1993.

[5] A. Sankar and C.-H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE Trans. on Audio and Speech Processing*, Vol. 4, No. 3, 1996.