Multipass algorithm for acquisition of salient acoustic morphemes

M. Levit[1]*, A. L. Gorin and J. H. Wright*

AT&T Laboratories-Research
180 Park Ave., Florham Park, New Jersey 07932, USA
{levit,algor,jwright}@research.att.com

Abstract

We are interested in spoken language understanding within the domain of automated telecommunication services. Our current methodology involves training statistical language models from large annotated corpora for recognition and understanding. Since the transcribing of large speech corpora is a resource consuming task, we are motivated to exploit speech *without* transcriptions. In particular, we learn the semantic associations for a task exploiting only phone-based sequences from the output of a task-independent ASR-system. In this paper we present a new multipass algorithm for acquiring salient phone sequences from untranscribed speech corpora and evaluate their utility for the HMIHY task. Compared to our previous strategy, this algorithm is shown to produce improved call-classification results while reducing up to 7-fold the number of salient phone-sequences selected for training.

1. Introduction

The subject of our research is machine understanding of spoken natural language. Our current methodology comprises word-based training which needs an annotated training corpus at the word level. Since the annotating of large amounts of speech data is time consuming and expensive, we are exploring the possibility of an understanding system that acquires lexicon, syntax and semantics from untranscribed speech. In particular, our strategy makes use of clusters of semantically meaningful phone sequences, which we call *acoustic morphemes*, for classifying of utterances. The representations of the utterances at the phone level are obtained as an output of a task-independent phone recognizer [6].

We evaluate our algorithms for the *How May I Help You* (HMIHY) task [2], where an automated dialogue system is designed to infer an appropriate machine action upon the service requests made over the phone by non-expert users. The requests are made in form of natural language utterances and elicited by an open-end prompt *"How May I Help You?"*.

There are several differences between our methodology and methods for inferring of words from the subword sequences described in the literature (see e.g. [1]). First of all, we exploit the semantic significance (*salience*) of phone sequences (*phrases*), generalizing thus the problem of learning from speech alone to learning from speech and meaning. The utility of semantics for learning to understand language has been proven in [3, 7]. The second feature incorporated in our strategy allows us to handle the output of an imperfect ASR by combining similar phrases into clusters. Finally, there are no restrictions on the length of the extracted sequences.

In this paper we propose a new algorithm for acquisition of salient phone phrases from training data. For this purpose we employ an iterative scheme. Every iteration includes extracting of phone phrases from the training corpus and parsing this corpus represented at the phone level with these phrases. Using a ML-parser distinguishes the presented multipass algorithm from our previously described two-pass strategy which employed a simple filter [4, 5].

The rest of the paper is organized as follows: in the next section we present a short overview of the data set we are using in our experiments. Section 3 addresses acoustic morphemes. The iterative procedure for extracting of the salient phone phrases from the training corpus is described in Section 4. Results of its application are presented in Section 5, and the conclusion is given in Section 6.

2. Database

Our database is a collection of sentences generated from the recordings of callers responding to the prompt *"AT&T. How may I help you?"* [2]. There are 7642 and 1000 sentences in our training and test sets respectively. Sentences are represented at the phone level and provided with semantic labels drawn from 15 call-types including an open-class denoted "OTHER". Phone lattices are produced by a task independent phone recognizer [6]. The best-path ASR-output is denoted *ASR-phone*. For baseline comparison purposes we also consider transcriptions at the phone level obtained by replacing every word in the word level annotations by its most likely dictionary pronunciation. This data set is called *transcr-phone*.

3. Acoustic Morphemes

In our methodology we use semantic associations of selected sequences of phones to classify the whole utterance whose part they are. To be selected the phone sequence (*phrase*) $f = [p_1 p_2 \dots p_k]$ must be meaningful and entropy reducing. Selected phrases are then combined into clusters (*acoustic morphemes*) based on acoustic and semantic similarity measures. These clusters are then represented as *Finite State Machines*.

4. Algorithm

The scheme of the multipass algorithm for extracting salient phrases is shown in Figure 1. At this point we present a high-level description of the algorithm with specific details explained in the subsequent sections.

On each iteration we examine *phrases*: sequences of *events* which, on their part, either are phones (on the first iteration) or are created based upon phrases selected on the previous iteration. Denote by $C(0)$ the initial corpus represented at the phone level, we then iterate as follows:

Given: corpus $C(t-1)$ from the previous iteration — set of sentences represented as finite sequences of events;

Generator: create $F(t)$: set of phrases (subsequences of observed events) consisting of $\leq n$ events pruned based on entropy and salience criteria (Section 4.1);

Stop condition: there are no significant changes in $F(t)$ compared to the previous iteration (Section 4.4);

[1] Also: Friedrich-Alexander-University Erlangen-Nuremberg

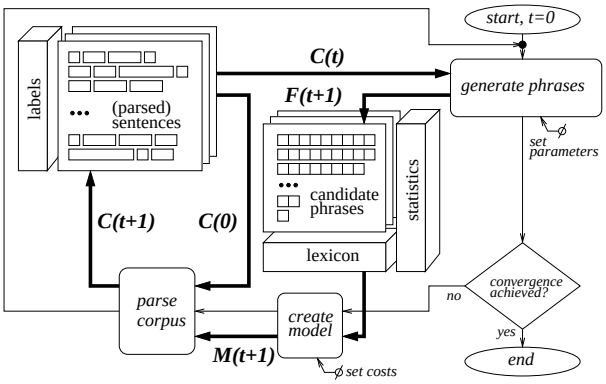

Figure 1: Iterative procedure for extracting salient phone phrases from the training set.

Model: create a stochastic language model $M(t)$ using the selected phrases from $F(t)$ as a lexicon (Section 4.2);

Parsing: parse the original corpus using this model and express it in terms of the phrases from $F(t)$, creating the corpus $C(t)$ (Section 4.3);

Loop variable: define the new set of events as the phrases from $F(t)$ observed in $C(t)$ plus the 51 phones.

To illustrate this procedure consider phone sequence from $C(0)$:

ay n iy D T uw m ey K ey K ax l eh K T K ao l

which represents the sentence *I need to make a collect call*. Let $n = 4$. Under our experiment conditions the set of phrases selected on the first iteration includes phrases [*n iy D*], [*m ey K ey*], [*K ax l*] and [*T K ao l*], and the parser segments the original sequence into the sequence from $C(1)$:

ay n_iy_D T uw m_ey_K_ey K_ax_l eh K T_K_ao_l,

where *a_b_c* denotes a new event representing the phrase [*a b c*]. On the 2^{nd} iteration we acquire new longer phrases: [*K_ax_l eh K T_K_ao_l*], [*n_iy_D T uw m_ey_K_ey*] and [*ay n_iy_D T uw*], so that into $C(2)$ goes the sequence:

ay n_iy_D_T_uw_m_ey_K_ey K_ax_l_eh_K_T_K_ao_l.

4.1. Phrase generator

We now describe the generator module of the algorithm in more detail. On each iteration, we prune the set of observed phrases based on three criteria. Given a phrase $f = [p_1 p_2 \dots p_k]$ we compute its:

- utility for within-language modeling (reducing entropy): In particular, define the mutual information (MI) of f as:

$$I(f) = \log_2 \frac{P(p_1 \dots p_k)}{P(p_1) \dots P(p_k)}$$

  and $I^{norm}(f) = I(f)/\text{length}_{phone}(f)$, where $\text{length}_{phone}(f)$ is the number of phones comprised in the phrase f (k);

- utility for understanding (salience for the task): In particular, a simple salience measure is:

$$P_{max}(f) = \max_c Pr(c|f).$$

  Also a more general measure based on Kullback-Leibler distance can be used [3];

- reliability of these characteristics: The number of occurances $\#f$ of the phrase f in the corpus is a simple correlate of reliability. One more precise criterion of reliability of the salience measure is given by the *multinomial significance test* [9]. It examines possible partitions of total of $\#f$ observations of phrase f in different semantic classes. The probabilities of the partitions are then estimated under the null-hypotheses of the statistical independence of f and c. If the sum of probabilities of partitions which are less probable than the actually observed one is less than some threshold α, phrase f is accepted.

4.2. Creating the language model

On each iteration probability of any phrase is *Maximum-Likelihood*-estimated within the set of phrases consisting of the same number of events: $p(f) = \#f / \sum_i \#f_i$, $\text{length}(f) = \text{length}(f_i)\ \forall i$. On the iteration to follow the phrases of different lengths which survived pruning and occur in the new representation of the corpus will be represented as events (and thus as phrases consisting of one event) themselves, so that their probabilities will be re-estimated and normalized within a set they all belong to. As we proceed with the iterations and the process converges (only few phrases of more than one events are generated – see next section) we finally obtain a unigram stochastic language model containing all selected phrases.

4.3. Parser

In our experiments we used *Finite State Machines* to perform a ML-parsing of the corpus. Therefore every phrase is provided with the scores equal to its mutual information (without length normalization). Among all possible competing segmentations the one with the highest sum of scores is then chosen. It is not difficult to see that the parser built this way is equivalent to a ML-parser.

4.4. Remarks on convergence

The reason why the iterative process converges is its relation to the EM algorithm with the phone-representations of the sentences as observed variables and their segmentations in phrases as hidden variables. However, since we Viterbi-re-estimate statistics (based on the best path returned by the ML-parser) it is not the classical version of EM-algorithm but the so called EM* simplification [8]. In fact it takes only a few iterations before the process converges. Here we say that convergence is attained if the number of one-event-phrases selected on some iteration doesn't exceed 5% of the total number of phrases selected on this iteration.

5. Experiments

For our experiments the following strategy turned out to be the most successful: we divide iterations in two phases. During the first phase we reduce the entropy of the corpus: the generator module only selects the phrases which occur frequently enough in the corpus and possess relatively high values of mutual information (being thus important for the within-language modeling task). In our experiments we reduced the normalized entropy from 5.0 to 2.3 bit/phone.

Once the convergence is attained we introduce the salience threshold $P_{max}(f)$ combined with the multinomial significance test in the generator module reducing thus the set of the phrases common in the language down to the phrases which are also charged with strong semantic associations. This strategy is justified by the observation that shorter phone sequences (unlike shorter word sequences) are of low salience [4].

The thresholds for the generator module were set:

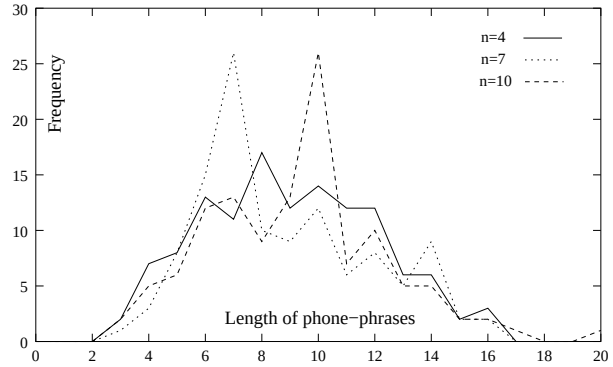

Figure 2: Length of selected phrases after introducing of the salience threshold and multinomial significance test.

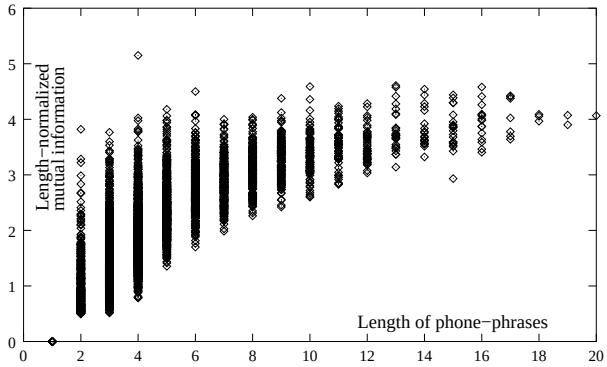

Figure 3: Normalized mutual information of phone phrases at the end of the first phase of the iterations; $n = 4$.

- mutual information: $I^{norm}(f) \geq 0.5$;
- salience: $P_{max}(f) \geq 0.5$;
- number of occurances: $\#f \geq 5$;
- multinomial significance threshold $\alpha = 0.05$ (see [9] for details).

We conducted our experiments using 7462 training and 1000 test utterances, each set labeled in two different ways as described in Section 2. After the classifier [9] has been trained with selected acoustic morphemes we apply it to the test utterances to classify them as one of 15 call-types or reject.

5.1. Results

Our first experiment on *ASR-phone* determined the optimal value for the parameter n: the maximal length of phrases in events considered on every iteration.

Figure 2 shows for different n the distributions of salient entropy reducing phrases selected by the algorithm over the number of phones they consist of. The reason for the differences in shapes the curves exhibit is that longer phrases tend to possess higher values of salience and mutual information [4]. This leads to the algorithmical artifacts: maxima by n and its multiples, which can be clearly seen in the distributions $n = 7, 10$, whereas the distribution $n = 4$ exhibits a rather smooth shape which also seems to be more natural.

In fact the longer phrases tend to have higher values of the length-normalized MI too. The dependency between $\text{length}_{phone}(f)$ and $I^{norm}(f)$ at the end of the entropy reducing phase of iterative process is plotted in Figure 3.

The number of selected salient phrases after convergence, the number of iterations and elapsed CPU-time are given for different n in the following table:

experiment	iter.	selected phrases		time
		entr. red.	sal.+sign.test	
$n = 2$	6	5170	217	15 min
$n = 4$	6	5168	239	20 min
$n = 7$	5	5230	232	35 min
$n = 10$	5	5304	237	55 min

We observe that the iterative process described in Section 4 has a clearly better time behavior for the smaller values of n. Compared to the two-pass algorithm with filter (2P-F) [4] which produced 1691 salient phrases, the multipass algorithms result in a 7-fold reduction of salient phone phrases.

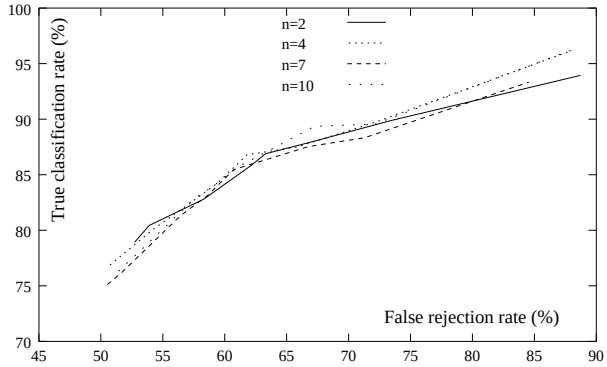

Figure 4: Influence of the maximal phrase length on every iteration (n) on the call-classification performance on speech.

To assess the impact of n on the classifier performance we employ another evaluation criterion: the ROC-curve which reflects dependency of the True Classification Rate on the False Rejection Rate, varying the rejection salience-threshold for the classification[2].

From Figure 4 we see that the choice of parameter n is not decisive for the performance of the classifier. We also trained the classifier on the union of phrases selected by three processes ($n = 4, 7, 10$), increased their number up to 363, which yielded only a 5%-extension of the ROC-curve in the direction of the lower False Rejection Rates (Figure 5). We conclude thus that differences in the sets of selected phrases we obtain for different parameter n don't affect the performance. Compared to 2P-F algorithm [4] we achieved slight improvements of true recognition rates while the working area slid by 10% in the direction of the higher False Rejection Rates (Figure 5). Higher FRR are explained by the fact that the most false rejections in the 2P-F algorithm were caused by not finding in the test utterances any acoustic morphemes at all, and the multipass algorithm reduced the number of selected phrases farther by factor seven.

Finally we classified on pruned lattices instead of best paths (reducing thus the FRR, [5]), our new strategy incorporated one additional improvement which allowed us to use for classification all cluster detections made in the pruned lattices of the test utterances, weighted by the probabilities of the lattice paths they lie on. In the previous version of the classification on lat-

[2]We focus here on the rank one results, in contrast to the rank two results in previous papers.

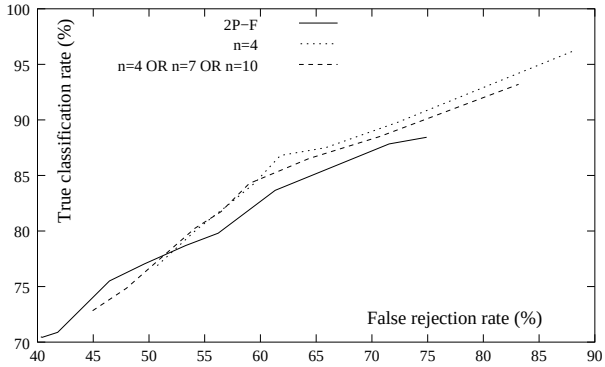

Figure 5: Comparison of classification results for the 2P-F algorithm, multipass algorithm with $n = 4$ and disjunction of multipass algorithms $n = 4, 7, 10$.

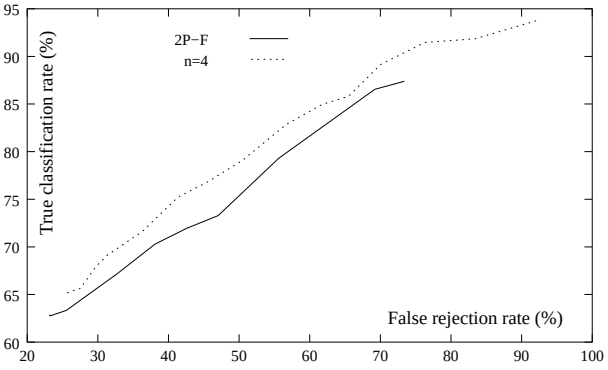

Figure 6: Comparison of classification results on pruned lattices for the 2P-F algorithm (only the most probable detections considered) and multipass algorithm with $n = 4$.

tices [5] only the detections made on the most probable path containing any detections at all were considered. The new algorithm resulted in 2-3% better ROC-curves (see Figure 6) with the 3-fold reduction in the number of selected salient phrases (multinomial significance test was not employed).

With the baseline experiment, we carried out on the *transcr-phone* corpus, we proved that the multipass algorithm can make the phone-based understanding system even outperform the word-based systems. For this purpose we compared the ROC-curves obtained on *transcr-phone* using the training algorithm presented above with the ROC-curve on the same corpus but represented at the word level and trained after one-pass training scheme. In both cases parameter setting $n = 4$ was used. The comparison in Figure 7 shows that automated phone-based understanding produces results even slightly better than the word-based understanding while requiring much less training expenses.

6. Conclusions

Our experiments show that the problem of training automated language understanding can be attacked at the phone level, saving the considerable effort of transcribing large amounts of training data. We described a new multipass algorithm for acquisition of salient phone phrases from untranscribed speech corpora. This algorithm is shown to reduce the number of extracted phrases by a factor of seven while producing results similar to our previous algorithm [4, 5]. We also obtained an

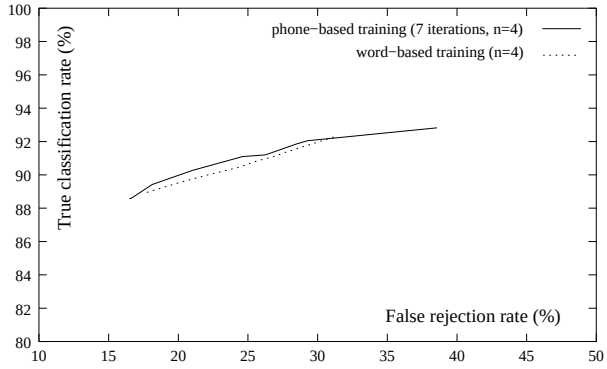

Figure 7: Call-classification performance on text with word-based standard and phone-based multipass strategies.

improvement of the ROC-curves by 2 percentage points with a 3-fold reduction. The best performance-to-time relation was achieved considering sequences of up to four events on every iteration when splitting the iterative process in two phases: find entropy reducing phone phrases and select those of them which are reliably salient.

7. References

[1] Deligne S. and Bimbot F.: *Inference of Variable-length Linguistic and Acoustic Units by Multigrams*. Speech Communication 23, pp. 223–241, 1997.

[2] Gorin A. L., Riccardi G. and Wright J. H.: *How may I help you?*. Speech Communication 23, pp. 113–127, 1997.

[3] Gorin A. L.: *On Automated Language Acquisition*. Journal of the Acoustical Society of America (JASA), 97(6), pp. 3441–3461, 1995.

[4] Gorin A. L., Petrovska-Delacrétaz D., Riccardi G. and Wright J. H.: *Learning Spoken Language without Transcriptions.* IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU'99, Colorado, USA, Dec. 1999.

[5] Petrovska-Delacrétaz D., Gorin A. L., Wright J. H. and Riccardi G.: *Detecting Acoustic Morphemes in Lattices for Spoken Language Understanding*. 6th International Conference on Spoken Language Processing, ICSLP'2000, Beijing, China, Oct. 2000.

[6] Riccardi G.: *On-line Learning of Acoustic and Lexical Units for Domain-Independent ASR.* 6th International Conference on Spoken Language Processing, ICSLP'2000, Beijing, China, Oct. 2000.

[7] Roy D.: *Learning Words from Sights and Sounds: A Computational Model* Ph.D. Thesis, MIT, 1999

[8] Schukat-Talamazzini E. G.: *Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen* Vieweg, Braunschweig, 1995

[9] Wright J. H., Gorin A. L. and Riccardi G.: *Automatic Acquisition of Salient Grammar Fragments for Call-Type Classification*. Proc. Eurospeech, Rhodes, Greece, Sept. 1997.