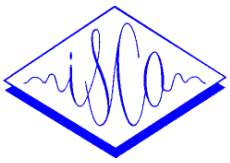

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

DETECTING *ACOUSTIC MORPHEMES* IN LATTICES FOR SPOKEN LANGUAGE UNDERSTANDING

D. Petrovska-Delacrétaz, A. L. Gorin, J. H. Wright and G. Riccardi

AT&T Labs, Speech Research, 180 Park Avenue, Flohram Park, N.J. 07932
{dijana, algor, jwright, dsp3}@research.att.com

ABSTRACT

Current methods for training statistical language models for recognition and understanding require large annotated corpora. The collection, transcription and labeling of such corpora is a major bottleneck for creating new applications and for refinements of existing ones. Thus, it is of great interest to develop methods for automatically learning vocabulary, grammar and semantics from a speech corpus *without* transcriptions. In this paper we report on an experiment where *acoustic morphemes* are automatically acquired from the output of a task-independent phone recognizer. The utility of these units is experimentally evaluated for call-type classification in the *'How may I help you?'* task. Detected occurrences of the acoustic morphemes in the lattice output provide the basis for the classification of the test sentences. Using lattices, we achieve a reduction of 59% from the false rejection rate using best paths, albeit with a 5% reduction in the correct classification performance from that baseline.

Keywords: Spoken language understanding, Salient phrase acquisition, Acoustic morphemes, Phone lattices.

1. INTRODUCTION

Our current methods for training spoken dialog systems require large annotated corpora. The speech is transcribed by human listeners and each utterance is semantically labeled. The resultant database is exploited to train stochastic language models for recognition [1] and understanding [2]. This database generation process is a major bottleneck for creating new applications and for refinements of existing ones. Thus, we are interested to develop methods for training spoken dialog systems based on speech *without* transcriptions.

In this paper we describe an experiment in that direction where we acquire automatically the vocabulary, grammar and semantics from a speech corpus without transcriptions. During the training phase *salient phone-phrases* are automatically acquired from the output of a task-independent phone recognizer using the methods of [3]. These phrases are furthermore clustered into *acoustic morphemes*, using the techniques of [2]. The utility of these units is experimentally evaluated for call-type classification in the *'How may I help you?'* (HMIHY) task [3]. Detected occurrences of the acoustic morphemes in the lattice output provide the basis for classification of the test sentences. This work extends our previous results [4], where we used only the best hypothesis of the phone recognizer.

There are several reports in the literature in this general direction. The earliest work is [5], demonstrating automatic acquisition of words and grammar from collapsed text. However, that work did not address the issues arising from non-perfect recognition of speech. In [6] it was shown how to acquire lexical units from speech alone without transcriptions and exploit them for spoken language understanding. That experiment was constrained to speech comprising isolated word sequences and used Dynamic Time Warping (DTW) matching to decide if an observation was a new word or variation of a known word. Several recent experiments [7] – [10] report attempts to acquire variable-length units from speech alone using sub-word methods.

While one can learn much about a spoken language by merely listening to it, one can progress further and faster by exploiting semantics. This has been demonstrated in both engineering domain [11] and in analysis of children's language acquisition [12]. In this research, we exploit speech plus meaning for learning to understand without transcriptions. The meaning in our case comprises the semantic labels (actions) associated to the utterances. The actions can be extracted automatically from either wizard experiments [13] or from autonomous dialogs [11].

The outline of this paper is the following: Section 2 describes the database that underlies the experiment. The automated acquisition procedure of the acoustic morphemes is summarized in Section 3. Section 4 introduces the general ideas about lattices and their utility for improving coverage of the test sentences. The experimental evaluation of these acquired units for call-type classification is reported in Section 5. The conclusions are given in Section 6.

2. DATABASE

The database used for the experiments is generated from recordings of users talking to human agents over the telephone, responding to the prompt *"AT&T. How may I help you?"*. The characteristics of this data and early experiments can be found in [3]. These human/human speech transactions, with their related call-labels are separated into a training and testing sets of 7462 and 1000 spoken utterances. They are denoted as *train* and *test*.

In this work, the utterances are processed through a task-independent phone recognizer. In particular, a phonotactic language model is trained on the *Switchboard 1* corpus [14], using a Variable-length N-gram Stochastic Automaton (VNSA) [15]. This corpus is unrelated to the HMIHY task, except in that they both comprise fluent English speech. Off-the-shelf telephony acoustic models are used. The phone accuracy of this recognizer on the HMIHY test speech data is 44% and the phone lattice accuracy 68%. The training and the test sets so generated are denoted by *ASR-phone-train* and *test*. Their mean length is 54 phones per utterance [4]. The classification of the test utterances is achieved by exploiting the lattice output of the ASR

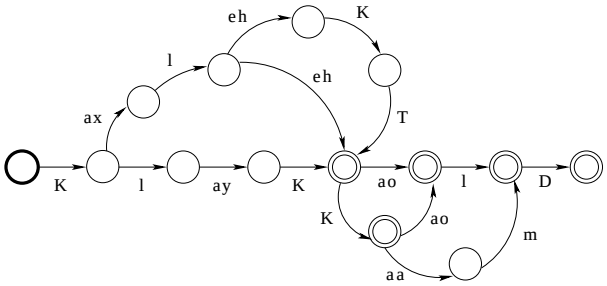

Figure 1: Example of an acoustic morpheme, associated to the call-type *collect*.

phone recognizer, denoted as *ASR-phone-latt-test*.

For baseline comparisons we will also use the orthographic transcriptions of the database. They are denoted as *transcr-word-train* and *test*. We also generated 'noiseless' phonetic transcriptions from the *transcr-word* data by replacing each word by its most likely dictionary pronunciation and omitting word delimiters. We denote these data sets as *transcr-phone-train* and *test*.

3. ACOUSTIC MORPHEMES

Acoustic morphemes are automatically acquired from the output of a task-independent phone recognizer. First, *candidate phone-phrases* are acquired by searching the space of observed phone sequences from *ASR-phone-train* using the algorithm of [4]. Consider a candidate phone-phrase f. A simplified measure of its salience [11] for the task is the maximum of the *a posteriori distribution*: $P_{max}(f) = \max_{c_i \in C} P(c_i|f)$, where C refers to the 15 call-types from the HMIHY task [3]. The goal is to select a subset of the candidate phone-phrases, denoted as *salient phone-phrases*. This selection is done by applying a threshold on $P_{max} \geq 0.6$ and by using a multinomial statistical significance test, as described in [2]. This significance test excludes low-frequency phrases for which a fortunate conjunction of events can give a high apparent salience purely by chance. We test the hypothesis that the observed call-type count distribution is a random sample from the prior distribution.

The salient phrases are clustered before they are used for classification [2]. The clustering is achieved using a combination of string and semantic distortion measures. Each cluster is then compactly represented as a finite state machine (FSM). These clustering methods are applied to the set of 1688 salient phone-phrases, leading to 470 *acoustic morphemes*. An example of a subgraph of one of those morphemes, related to the call-type *collect*, is given in Figure 1, where a bold circle represents an initial state and a double circle a final state.

4. IMPROVED COVERAGE WITH LATTICES

4.1. Coverage of test utterances with acoustic morphemes

The classification is based on detected occurrences of acoustic morphemes within the test utterances. Working with the output of the non-perfect ASR phone recognizer introduces the problem of the coverage of the test sentences with the acoustic morphemes. We show how exploiting the lattice output of the task-independent phone recognizer improves the coverage.

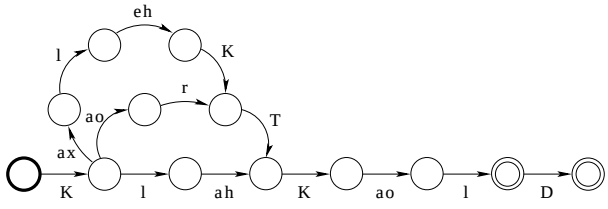

Figure 2: Phone lattice (from the ASR phone recognizer output) for the utterance *"collect call"*.

Experiment	*No detections*
best path	42%
pruned	12%
full lattice	6%

Table 1: Comparison of the percentage of test sentences with *no* detected acoustic morphemes, in the best path, in the pruned lattice and in the full lattice experiments.

Lattices are efficient representations of a distribution of alternative hypothesis [16] [17]. To fix our ideas, a simple example of a lattice network, resulting from the utterance *"collect call"*, is shown in Figure 2, where a bold circle represents an initial state and a double circle a final state. The most likely phone sequence of the transcribed utterance is *"K ax l eh K T K ao l"*, and the best path of the ASR phone recognizer is *"K l ah K ao l"*. Whereas the salient phone-phrase *"K ax l eh K T K ao l"* is not present in the best path, it does appear in the lattice network. Exploiting lattices results in additional matches of the salient phrases in the utterances, as compared to searching only in the best paths.

In speech recognition the weights (likelihoods) of the paths of the lattices are interpreted as negative logarithms of the probabilities. For practical purposes, we consider also the case of the pruned network. In this case we restrict the beam search in the lattice output, by considering only the paths with probabilities above a certain threshold relative to the best path. The threshold r is defined as: $r_i \leq r$, with $r_i = p_i \,/\, p_1$, where p_i is the probability of the i^{th} path and p_1 is the probability of the best path.

In order to quantify the coverage, we first measured the number of test sentences with *no* detected occurences of acoustic morphemes, for the experiments using best paths, pruned lattices and full lattices. These numbers are reported in Table 1. Observe that 42% of the best path sentences have no detected acoustic morphemes. When we expand the search of the acoustic morphemes to the pruned lattices, the number of sentences with no detections decreases to 12%. This number drops down to 6% when we search in the full lattices.

The relative frequency distributions of the number of detected acoustic morphemes in the best paths, in the pruned lattices, and in the full network experiments, are shown in Figure 3. As expected, the number of detections increases in the experiments using lattices.

4.2. Statistics of a particular Acoustic Morpheme F_c

In the previous section we concluded that expanding the search of the acoustic morphemes in the lattice network results in im-

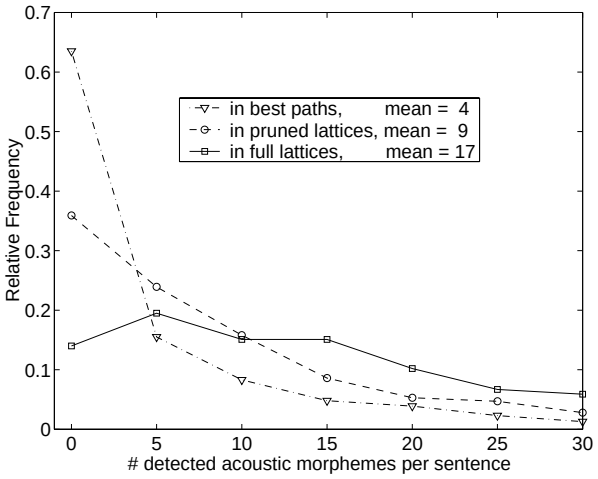

Figure 3: Relative frequency distribution of the number of detected acoustic morphemes per test sentence, measured on the best path, on the pruned lattice, and on the full lattice experiments.

proved coverage of the test sentences. It is of course accompanied by an increased number of false detections of the acoustic morphemes. In order to study in more detail the false detection issue, we will focus on one particular morpheme F_c, strongly associated to the call-type *c = collect*. We will compare its occurrences in the best paths and in the lattice network. The salient phone-phrases clustered in this morpheme represent variations of the phrase *"collect call"*. A subgraph of this morpheme was shown in Figure 1. Its salience on the training set is $P(c|F_c) = 0.89$. We introduce the following notation: W = {manually chosen word sequences corresponding to F_c}.

Table 2 shows the comparison of the coverage of the test utterances with the acoustic morpheme F_c, on the best paths and on the lattice network. F_c is detected in 3% of the best paths. Searching in the full lattice network, increases this coverage to 8%. However, in *transcr-word-test*, the word sequences corresponding to the Acoustic Morpheme F_c are present in only 7% of the transcribed sentences.

lattice	*r*	$P(F_c)$	$P(W)$
best path	1.00	0.028	0.071
pruned	0.05	0.042	-
full lattice	0.00	0.080	-

Table 2: Coverage of the test utterances with the acoustic morpheme F_c; $P(W)$ is calculated on the *transcr-word-test* set.

Table 3 illustrates the relationship between the detections of F_c given the call-type, $P(F_c|c)$, and its salience, $P(c|F_c)$, measured on the test set. Observe that the number of detections of this morpheme given the call-type increases from 15% in the case of the best paths, to 31% in the case of the full lattice search. In parallel, we observe a decrease in the salience from 93% to 71%. This is an indication that the high salience of this morpheme on the best paths is conserved in the pruned lattices, but not in the full lattices.

lattice	*r*	$P(F_c|c)$	$P(c|F_c)$
best path	1.00	0.15	0.93
pruned	0.05	0.20	0.90
full lattice	0.00	0.31	0.71

Table 3: Detections of F_c given the call-type *collect*, and its salience.

Table 4 illustrates the 'recognition accuracy' of the Acoustic Morpheme F_c, as compared to the transcribed text. The probabilities $P(F_c|W)$ and $P(F_c|\bar{W})$ indicate how often do we find the acoustic morpheme in the ASR phone output, given that we know that the corresponding word sequences are present (or not) in the transcribed sentences. Searching in the full lattice increases the number of detected occurrences of F_c from 38% up to 66%, albeit with a parallel increase of the falsely detected morphemes.

Lattice	*r*	$P(F_c|W)$	$P(F_c|\bar{W})$
best path	1.00	0.38	0.001
pruned	0.05	0.53	0.004
full lattice	0.00	0.66	0.035

Table 4: 'Recognition accuracy' of the Acoustic Morpheme F_c, as compared to the transcribed data.

5. EXPERIMENTAL EVALUATION WITH A CALL-TYPE CLASSIFIER

5.1. Experimental setup

The utility of the *acoustic morphemes* is evaluated for call-type classification in the HMIHY task following the methodology of [3]. We classify the test utterances by detecting occurrences of acoustic morphemes. To review, an utterance is classified by the system as one of the 14 call-types or rejected as *'other'*. Rejection is based on a salience threshold for the resulting classification [2]. One dimension of performance is the *False Rejection Rate* (FRR), which is the probability that an utterance is rejected in the case that the user wanted one of the call-types. The cost of such an error is a lost opportunity for automation. The second dimension of performance is the *Probability of Correct Classification* (P_c), when the machine attempts a decision. The cost of such an error is that of recovery via dialog. Varying the rejection threshold traces a performance curve with axes P_c and FRR.

Searching in the lattice network will introduce the additional problem of multiple detections of the acoustic morphemes on different levels of the lattice network, and the issue of combining them optimally. For an evaluation of the usefulness of the phone lattices, we will defer the problem of treating multiple detections from different levels of the network, and stop our search in the lattice network as soon as we find a sentence with one or more detections. We modified the existing call-type classifier in the following way: for the test sentences without detected occurrences of the acoustic morphemes in the best paths, we expand our search in the lattice network and stop as soon as we find a sentence with one or more detections.

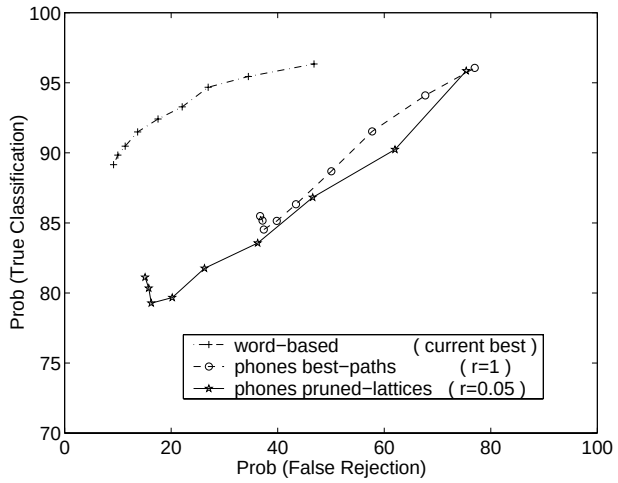

Figure 4: Call-classification performance using pruned test lattices compared to the single best hypothesis experiment. Our current best results using using words are also included.

5.2. Classification results

Our extensive studies of call-type classification experiments and numeric language recognition, using *salient grammar fragments* acquired from word transcribed text and performed on the same task and database, can be found in [1]–[3] [18]. In those experiments, we have reported on call-classification experiments trained on word transcriptions, and exploiting spoken language understanding with utterance verification. Those results are also included in Figure 4 for comparison purposes. The results reporting the utility of the detected occurrences of the acoustic morphemes in the pruned lattice test sentences are shown in Figure 4. Using this new lattice-based detection method, we achieve an operating point with 81% correct classification rate at rank 2, with 15% false rejection rate. This is a reduction of 59% from the previous false rejection rate using best paths, albeit with a 5% reduction in the correct classification performance from that baseline.

6. CONCLUSIONS

From the experiments reported in this paper we conclude that exploiting the lattice output of the ASR phone recognizer increases the number of test utterances with detected occurrences of acoustic morphemes from 58% in the case of the best paths, to 88% in the case of the pruned lattice network, to 94% in the full lattice. In the case of the pruned lattice search this is a relative improvement in coverage of 52% over the best path results.

Using this new lattice-based detection method for call-type classification, we achieve a useful operating point with 81% correct classification rate at rank 2, with 15% false rejection rate. This is a reduction of 59% from the previous false rejection rate using best paths, albeit with a 5% reduction in the correct classification performance from that baseline.

7. REFERENCES

[1] Riccardi G. and Gorin A. L.: *Spoken Language Adaptation over Time and State in Natural Spoken Dialog Systems*. IEEE Trans. on Speech and Audio Proc., Vol. 8, No. 1, Jan. 2000.

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

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

[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] Olivier D. C.: *Stochastic Grammars and Language Acquisition Mechanism*. Ph.D. Thesis, Harvard Univ., 1968.

[6] Gorin A. L., Levinson S. and Sankar A.: *An Experiment in Spoken Language Acquisition*. IEEE Trans. on Speech and Audio, pp. 224–240, vol 2, no 1, part II, Jan. 1994.

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

[8] Lloyd-Thomas H., Parris E. and Wright J. W.: *Recurrent Substrings and Data Fusion for Language Recognition*. Proc. ICSLP, Sydney, Australia, Dec. 1998.

[9] Herbeck S. and Ohler U.: *Multigrams for Language Identification*. Proc. Eurospeech, Budapest, Hungary, Sept. 1999.

[10] Chollet G., Černocký J., Constantinescu A., Deligne S. and Bimbot F.: *Towards ALISP: a proposal for Automatic Language Independent Speech Processing*. Computational Models of Speech Pattern Processing (ed. Ponting K.), NATO ASI Series, vol. 169, pp. 375–388, 1999.

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

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

[13] Ammicht E., Gorin A. L. and Alonso T.: *Knowledge Collection for Natural Language Spoken Dialog System*. Proc. Eurospeech, Budapest, Hungary, Sept. 1999.

[14] Linguistic Data Consortium, *http://morph.ldc.upenn.edu*.

[15] Riccardi G., Pieraccini R. and Bocchieri E.: *Stochastic Automata for Language Modeling*. Computer Speech and Language, vol 10(4), pp. 265–293, 1996.

[16] Llolje A., Pereira F. and Riley M.: *Efficient General Lattice Generation and Rescoring*, Proc. Eurospeech, Budapest, Hungary, vol 3, pp. 1251–1254, Sept. 1999.

[17] Mohri M. and Riley M.: *Network Optimization for Large-vocabulary Speech Recognition*, Speech Communication 28, pp. 1–12, Sept. 1999.

[18] Rahim M., Riccardi G., Saul L., Wright J. W., Buntschuh B. and Gorin A. L.: *Robust Numeric Recognition in Spoken Language Dialog*. To appear in Speech Communication.