MODELLING UNKNOWN WORDS IN SPONTANEOUS SPEECH

T. Kemp

Interactive Systems Laboratories
Department of Computer Science
University of Karlsruhe
76131 Karlsruhe, Germany

A. Jusek

AG Angewandte Informatik
Technische Fakultät
University of Bielefeld
Postfach 100131
33501 Bielefeld, Germany

ABSTRACT

In this paper we describe our experiments with different acoustic and language models for unknown words in spontaneous speech. We propose a syllable based approach for the acoustic modelling of new words. Several models of different degrees of complexity are evaluated against each other. We show that the modelling of new words can decrease the error rate in the recognition of spontaneous human-to-human speech. In addition, the new word models can be used as a measure of confidence capable of detecting errors in the recognition of spontaneous speech. Although the best performance is reached by applying phonetic a-priori knowledge in the design of the acoustic models, a pure data-driven approach is proposed which performs only slightly less efficiently.

INTRODUCTION

New words, i.e. words that are not in the recognition vocabulary, are an intrinsic problem to any real-world application of speech recognition. Each unknown word will inevitably lead to a recognition error. Therefore it would be desirable to have a detector for unknown words, so that appropriate action (e.g. a repair dialog) can be initiated.

Two different approaches to solve this problem have been proposed. In the first one [1], a generic word model, usually consisting of phonemes with some constraints in their possible succession, is used. This word model competes in the decoder with the dictionary words. If none of the dictionary words fits accurately to the spoken string, the generic word model will usually have a higher a-posteriori probability than any of the dictionary words and will be hypothesized. The second approach, which is used by many speech recognition systems designed for read speech, is to expand the dictionary by a large amount of additional entries and thus try to decrease the out-of-vocabulary (OOV) rate. This method is particularly suited for tasks where large corpora of text data are available for language modelling of the additional words. Its demands regarding memory and computational power are very high.

For read speech and vocabularies exceeding approximately 1,000 words, the extension of the dictionary usually outperforms the modelling of new words with generic new word models [2]. However, for spontaneous speech the situation is somewhat different, as up to 50% of the new words in spontaneous data are word fragments and cannot even be covered by a dictionary of arbitrary size. Recent experimental results on spontaneous data [3] seem to indicate that these word fragments can be adequately modelled with a generic new word model. This is particularly interesting as there are no large corpora of spontaneous speech available, that can be used for language modelling. Therefore, no reliable statistics for the words that are added to the dictionary can be derived anyway.

In this paper we describe our experiments with different generic word models used for spontaneous speech. We propose a syllable-based model which models the phoneme sequence constraints as a finite state automaton. A phonotactic approach using phonetic knowledge is compared to a pure data-driven approach. We also derive a statistical (bigram) language model that predicts probabilities for unknown words.

1. THE ACOUSTIC MODELLING

An acoustic model for new words must be capable to model all possible new words. A common approach to this problem is to allow *any* sequence of phonemes and use a bigram or trigram language model to capture the different probabilities of phoneme sequences. With proper smoothing of the language model, such a model has a 100% coverage of new words.

However, there is more structure in language that can be used to construct a new word model. Whereas there are more than $3 * 10^5$ words in a language, the number of *syllables* is one or two orders of magnitude lower, see table 1[1] . Syllable models thus offer a handy alternative to phoneme models. They offer tighter constraints to the possible sequence of phonemes and still do not require the high amount of resources that are needed when no generic models are used but instead the dictionary is expanded. In the following subsections, we introduce different syllable models.

1.1. The phonotactic model

The phonotactics of a language describes the combinatorial structure of syllables, words etc. A German syllable consists at least of a nucleus that can either be a vowel or a

[1] This data has been derived using the CELEX lexical database, ©MPI for Psycholinguistics, NL-6500 AH Nijmegen

base units	number of unique base units
words	315,694
syllables	10,817

Table 1. Number of words and syllables in German

diphtong. Consonant clusters can enclose the nucleus and must fulfil the phonotactic restrictions to form a valid syllable. For instance "spl" is a valid initial sound of a German syllable, while "lps" is not. We distinguish two types of syllables, reduced and non-reduced ones. The main difference is, that reduced syllables allow neutral vowels in the nucleus. According to the phonotactic rules of these two types of syllables we built a generic model for arbitrary German words based on the recognizers phonetic inventory [4]. This model was added to the recognizers dictionary.

1.2. The augmented phonotactic model

The described phonotactic model implicitely assigns unity probabilities to all state transitions between its phoneme states. As this may lead to insertions of phonetically possible while very infrequent transitions, we estimated transition probabilities for all transitions in the following way:
First, all words of the training corpus that appeared only once were extracted. These words were assumed to be good representatives of new words, as they occurred very infrequently in the training material. For each of these words, we took the phonetic transcription and computed the state sequence in the phonotactic automaton that corresponded to this phonetic transcription. All state-to-state transitions were counted in this way and the counts for each state transition were added together across the different words. After this step, the probabilities were estimated to be

$$p(A \rightarrow x) = \frac{c(A \rightarrow x)}{\sum_{n=1}^{N} c(A \rightarrow succ_n(A))} \qquad (1)$$

where $c(A \rightarrow x)$ denotes the number of times the transition from state A to state x was used. $succ_n(A)$ is the n-th possible successor of state A and N the number of possible successor states to A.

1.3. The data-driven model

The two models described so far make use of phonetic knowledge. If no such knowledge is available, e.g. if a speech recognition system for a new language has to be designed, it would be advantageous to have a model that can be constructed with less a-priori knowledge. Therefore we evaluated a data-driven model that derives the syllable model from a dictionary. We used the CELEX dictionary that contained 359,611 different words and, after stripping most of the foreign words and the stress markers, 10,671 different syllables. These syllables had 8,983,352 occurrences in the sample text used by CELEX. We constructed a minimal graph that encoded all syllables. This graph had 5485 states and 15101 transitions where the states were associated with the phonemes. Then we parsed as described above all 8,983,352 occurrences of the syllables, counted the state transitions and computed state transition probabilities using (1). Experimental results with this model are denoted 'data-driven 2' in table 2.
To find out the effects of the domain of the data, we derived another model using a dictionary taken from the scheduling domain [1] and trained the state transitions of this model using the same 1982-word list that was used to train the augmented phonotactic model. The results with this model are denoted 'data-driven 1' in table 2.

2. LANGUAGE MODELLING OF NEW WORDS

The different new word models were incorporated in the JANUS-2 system by defining all states of the automata as ordinary dictionary words. The allowed transitions were added to the backoff bigram language model that was used by the recognizer. Under the assumption that a new word is equally likely to occur after any of the regular words, the monogram probability of the first state of the automaton is the only free parameter which has to be determined on a crossvalidation data set. All experiments carried out with this uniform probability distribution for new words are shown in table 2.
However, the probability of an unknown word depends on the identity of its predecessor. For example, after 'My name is' it is very likely to observe an unknown word (usually a name), while after 'I would like' a new word is unlikely simply because in most cases a 'to' will be observed. To improve the language model with respect to the prediction of unknown words, we extracted all words in the training corpus that occurred only once, doubled all observed transitions that contained one of these words and substituted the word by the <UNKNOWN> token. The test set perplexity decreased by 3.5% with respect to the uniform distribution for the unknown words.

2.1. Optimizing the language model

In [3] was observed that the entropy of the probability distribution *into* new words is lower than the entropy of the distribution *from* new words. This means, that it is easier to predict a new word from a seen vocabulary word, than to predict a vocabulary word with the knowledge that the last token was a new word. As we do not have large amounts of training data, the estimation of the probability distribution $p(unknown \rightarrow vocabword)$ is probably very inaccurate. So we tried replacing this probability distribution by backing off to the monogram probability distribution of the vocabulary words, which can be estimated more robustly. Doing so reduced the test set perplexity in different experiments by another 0.5% to 1%.

2.2. Databases

For all described experiments we used the GSST database, which has been collected simultaneously at four different sites under the VERBMOBIL[2] project. It consists of human-to-human spontaneous German dialogs in the appointment scheduling domain, i.e. two persons try to schedule a meeting within the next month. The data is sampled

[1] This dictionary was kindly provided by Ralf Kompe, University of Erlangen

[2] The VERBMOBIL project aims at the development of a large speech-to-speech translation system and is funded by the german ministry for science and technology (BMBF)

with 16 bit at 16 kHz in a quiet office environment using a close-speaking microphone.

The database contains about 183,000 words of speech and has a bigram test set perplexity of around 95.

2.3. The JANUS-2 system

The speech-to-speech translation system JANUS-2 [6] [7] is a joint effort of the Interactive Systems Labs at Carnegie Mellon University, Pittsburgh, and at the University of Karlsruhe, Germany.
The baseline speech recognition component of JANUS-2 uses mixture-gaussian densities with a scalable amount of parameter tying. For the experiments described, we used 1677 decision-tree clustered context-dependent subtriphones which shared 1338 codebooks. In the preprocessing stage mel-scale spectra with a frame rate of 10 ms and their deltas, power, zerocrossing rate and peak-to-peak value were computed. The 37-dimensional input vector was transformed by linear discriminant analysis (LDA [8]) and split into two 16-component data streams. Training was done with Viterbi alignment. To capture some of the effects of spontaneous speech, specialized noise nodels were included [10].

The decoder computes word lattices with a multi-pass strategy. Trigrams and cross-word triphones can be employed, however, in the experiments described in this paper no cross-word models were used. After recognition, a maximum likelihood codebook adaptation using the recognition result and an additional recognition run are performed. For a more detailed description, refer to [5] [11] [12] [9].

The JANUS-2 decoder achieved a word error rate of 28.6% in the VERBMOBIL June 95 evaluation. This was the lowest error rate of the 5 participating institutions. For reasons of efficiency, in the experiments described we skipped the final adaptation step and reduced the number of parameters by 50% as compared to our evaluation system. The resulting system has a word error rate of 31.5%, which is our baseline performance.

3. RESULTS

For all described experiments, we used the 'short' subsection of the official VERBMOBIL June 95 evaluation. This testset contains 265 utterances with 3823 words. 122 (3.19%) of them are unknown. Table 2 shows the results of the recognition experiments.

System	WA (perc.corr.)	New word performance	
	(all words)	correct	false alarms
baseline	68.5% (72.1%)	-	-
phonotactic	68.9% (72.4%)	10.6%	11.5%
augmented	68.9% (72.5%)	18.8%	31.1%
data driven 1	68.9% (72.5%)	18.0%	28.6%
data driven 2	68.7% (72.4%)	10.6%	13.1%

Table 2. Results of recognition experiments

The column 'correct' shows the correct matches of the new word model, i.e. where a new word was correctly hypothesized. 'False alarms' shows the number of incorrect hypotheses of new words, including inserted new words, normalized by their total frequency. The correct rate is lower than the false alarm rate for each of the models, but the total word accuracy still improves: many of the erroneous newword hypotheses are at places where a recognition error occurs anyway.

System	WA (perc.corr.)	New word performance	
	(all words)	correct	false alarms
augmented	69.0% (72.6%)	23.8%	31.1%

Table 3. Result with unknown words in the language model

This is due to the fact, that in regions of high acoustic uncertainty none of the regular word models has a high probability and thus the new word model is more likely to be hypothesized than in regions of low acoustic confusability. The newword model can therefore be used to mark recognition errors.
To evaluate its performance as a measure of confidence, we performed another experiment. In this, we counted each newword hypothesis as correct that was hypothesized at a place where the recognizer outputs a wrong token when used without newword models. The results achieved in this way are shown in table 4.

System	Confidence measure performance	
	correctly spotted	false alarms
phonotactic	25	2
augmented	54	7
data driven 1	49	8
data driven 2	27	2
augmented w LM	57	10

Table 4. Results of confidence measure experiments

The best results (92% of the marked words are correctly marked) can be achieved with the model 'data driven 2'. However, the augmented phonotactic model offers twice as many detections with a correct rate of still 88%.

4. CONCLUSION

In this work we have shown that generic syllable based new word models increase recognition performance of spontaneous speech with a dictionary size of 3,000 words. The generation of the models is straightforward, and integration into existing recognition systems can be performed without changes to the system by simply modifying the statistical language model. The results can also be used as a measure of confidence, where the error rate within the marked words is only about 10% at a word accuracy of roughly 70%.
Although the explicit prediction of new words in the language model leads to a perplexity reduction of 4%, the gain in word accuracy is statistically insignificant. We expect this situation to change when larger corpora of spontaneous speech for more robust language modelling become available.

This is in contrast to the results of [3], where significant improvements could be achieved on word lattices. However, the June 1995 VERBMOBIL evaluation has shown that new word detection rates in word lattices may not be significant to results on first best hypotheses.

5. ACKNOWLEDGEMENTS

This research was partly funded by grant 413-4001-01IV101S3 from the German Ministry of Science and Technology (BMBF) as a part of the VERBMOBIL project. The JANUS project was supported in part by the Advanced Research Project Agency and the US Department of Defense. The views and conclusions contained in this document are those of the authors.
The author wishes to thank all members of the Interactive Systems Labs for useful discussions and active support.
Special thanks also to Ralf Kompe (University of Erlangen) and Pablo Fetter (Daimler-Benz AG).

REFERENCES

[1] B. Suhm, *Erkennung und Transkription Neuer Wörter in der Spracherkennung*, Diplomarbeit, University of Karlsruhe, Karlsruhe, Germany, April 1993 (in German)

[2] L. Chase, R. Rosenfeld, A. Hauptmann, M. Ravishankar, E. Thayer, P. Placeway, R. Weide, C. Lu, *Improvements in Language, Lexical, and Phonetic Modeling in Sphinx-II*, in Proc. ARPA Spoken Language Technology Workshop, Austin, TX, January 1995

[3] P. Fetter, F. Class, U. Haiber, A. Kaltenmeier, U. Kilian, P. Regel-Brietzmann, *Detection of unknown words in spontaneous speech*, Proc. EUROSPEECH 1995

[4] A. Jusek, H. Rautenstrauch, G. A. Fink, F. Kummert, G. Sagerer, J. Carson-Berndsen, and D. Gibbon, *Detektion unbekannter Wörter mit Hilfe phonotaktischer Modelle*, In W.G. Kropatsch and H. Bischof, editors, Mustererkennung 94, 16. DAGM-Symposium und 18. Workshop der ÖAGM Wien, pages 238–245. 1994 (in German)

[5] M. Woszczyna, N. Coccaro, A. Eisele, A. Lavie, A. McNair, T. Polzin, I. Rogina, C.P. Rose, T. Sloboda, M. Tomita, J. Tsutsumi, N. Aoki-Waibel, A. Waibel, W. Ward, *Recent Advances in Janus, a Speech-to-Speech Translation System*, Proc. EUROSPEECH 1993, pp. 1295-1298

[6] M. Woszczyna, N.Aoki-Waibel, F.D.Buo, N. Coccaro, K. Horigushi, T. Kemp, A. Lavie, A. McNair, T. Polzin, I. Rogina, C.P. Rose, T. Schultz, B. Suhm, M. Tomita, A. Waibel, *Janus 93: Towards Spontaneous Speech Translation*, Proc. ICASSP-94, pp. 345-349, April 1994

[7] P.Geutner, B.Suhm, F.D.Buo, T.Kemp, L.Mayfield, A.E.McNair, I.Rogina, T.Schultz, T.Sloboda, W.Ward, M.Woszczyna, A.Waibel. *Integrating different learning approaches into a multilingual spoken language translation system*, in Proc. of the IJCAI workshop on New Approaches to Learning for Natural Language Processing, pp 33 ff, Montreal, Canada, August 1995

[8] G. Yu, R.Schwartz: *Discriminant analysis and supervised vector quantization for continuous speech recognition*, Proc. ICASSP 1990, pp. 685 ff.

[9] I. Rogina and A. Waibel, *Learning State-Dependent Stream Weights for Multi-Codebook HMM Speech Recognition Systems*, Proc. ICASSP 1994

[10] T. Schultz and I. Rogina, *Acoustic and Language Modeling of Human and Nonhuman Noises for Human-to-Human Spontaneous Speech Recognition*, Proc. ICASSP 1995, vol 1, pp 293-296

[11] T. Kemp. *Data-driven codebook adaptation in phonetically tied SCHMMs*, in Proc. ICASSP-95, vol 1, pp 477ff, Detroit, May 1995

[12] P.Geutner. *Using Morphology towards better large-vocabulary speech recognition systems*, in Proc. ICASSP-95, vol 1, pp 445ff, Detroit, May 1995

MODELLING UNKNOWN WORDS IN SPONTANEOUS SPEECH

T. Kemp

Interactive Systems Laboratories
Department of Computer Science
University of Karlsruhe
76131 Karlsruhe, Germany

A. Jusek

AG Angewandte Informatik
Technische Fakultät
University of Bielefeld
Postfach 100131
33501 Bielefeld, Germany

In this paper we describe our experiments with different acoustic and language models for unknown words in spontaneous speech. We propose a syllable based approach for the acoustic modelling of new words. Several models of different degrees of complexity are evaluated against each other. We show that the modelling of new words can decrease the error rate in the recognition of spontaneous human-to-human speech. In addition, the new word models can be used as a measure of confidence capable of detecting errors in the recognition of spontaneous speech. Although the best performance is reached by applying phonetic a-priori knowledge in the design of the acoustic models, a pure data-driven approach is proposed which performs only slightly less efficiently.