AUTOMATICALLY GENERATED MODELS FOR UNKNOWN WORDS

A. Jusek, G. A. Fink, F. Kummert, and G. Sagerer

Technical Faculty
University of Bielefeld

ABSTRACT – Especially in recognition of spontaneous speech it is necessary to cope with the occurrence of unknown words. We present an approach to unknown word detection which is integrated into a standard HMM speech recognizer. From the context dependent sub-word units, e.g. triphones, that can be found in the training database a generic word model can be derived automatically using the context restrictions to form valid sequences of sub-word units. This generic word model combines automatically derived knowledge about the phonotactics of the language considered with the modelling quality of context dependent acoustic units. Detection of unknown words is achieved adding this model to the recognizer's lexicon. We present results of experiments carried out on a large German spontaneous speech recognition task.

INTRODUCTION

In the task of automatic speech recognition one major problem is, that only a certain limited vocabulary in the range of 1000 to 20000 words, is available to the recognizer. If a user utters a word, that is not among these known words, erroneous recognition results will be produced. To avoid this some approaches have been developed in the last years for the detection of unknown words. Usually, a generic word model built from arbitrary sequences of phonemes or generic acoustic models are used. Obviously, better results should be achieved if the sequence of phonemes is not arbitrary, but restricted according to e.g. phonotactic knowledge about the language considered. This assumption could be confirmed in our earlier approach (Jusek *et al.* 1995) when we applied phonotactic restrictions that were taken from results of linguistic research.

Nevertheless, all these approaches have the same disadvantage. For the modelling of known words context–dependent phoneme models are applied and for the modelling of unknown words either generic or context–independent phoneme models are used. Therefore, the latter models do not really represent unknown words but constitute an inadequate modelling of arbitrary words only. Even if special care is taken to balance between the different scoring methods the distinguishing between known and unknown words is a difficult task and extremely depends on the modelling of known words and on the given application.

In this paper, we present a new approach for the modelling of unknown words based on the same context-dependent sub-word units as are used for the modelling of known words. The context information establishes restrictions about valid sequences of sub-word units. These restrictions approximately reflect the phonotactic structure of the words in the corpus and can be derived automatically and consistently with the modelling of known words. In order to obtain a generic model that is capable of describing arbitrary words the context restriction are relaxed e.g. by forming generalized phoneme classes thus allowing a larger variety of valid sequences of sub-word units. This automatically created generic word model can then be used for the detection of unknown words when added to the recognizer's vocabulary.

In the following section a brief overview of related approaches to the detection of unknown words is given. Then some methods to improve unknown word modelling in general are summarized. In the fourth section our new approach is described. The results reported in the fifth section show that the automatically generated model can successfully be applied to the detection of unknown words in continuous spontaneous speech.

APPROACHES TO DETECTION AND MODELLING OF UNKNOWN WORDS

Since around 1990 the detection of unknown words is a well known problem in automatic speech recognition. Several approaches have been published to solve this problem. This section summarizes some of them.

Threshold based detection mechanisms

A simple way to distinguish between known and unknown words makes use of the acoustic score of the word models. It is based on the assumption that each known word scores worse in an area of an unknown word than in an area, where this word has been uttered. For this reason it is possible to select a threshold for a given scenario and to judge a word as known or unknown according to its acoustic score with respect to that threshold (Kai & Nakagawa 1994; Itou *et al.* 1992). This approach seems to be only useful for isolated word recognizers, because it is difficult to detect word boundaries in continuous speech. For that reason areas of unknown words may be tiled with small known words that fit well enough.

An improvement of the threshold method was introduced in (Hayamizu 1993). There the threshold was built by comparing the acoustic score of the usual vocabulary oriented recognizer with a phoneme oriented one. This technique ensures, that the threshold is always adapted to the actual quality of the speech signal. If the speech is e.g. disturbed by background noise both recognizers are influenced similarly.

Generic word models

Another approach to detecting unknown words is based on a generic word model. In this case a model is incorporated into the recognizer's vocabulary, that can reflect the structure of arbitrary words. There are several ways to achieve this. The most simple one is to build a generic acoustic model, that reflects all given phonemes simultaneously. Such models can be constructed by averaging the output probability densities of all given phoneme models. A generic word model can then be constructed as a sequence of those generic acoustic models.

It is obvious, that such a generic word model will reflect the acoustic structure of all words. It is also obvious, that the acoustic structure will not be reflected very well (Sakamoto & Matsunaga 1995).

A different method to construct generic acoustic models is reported in (Fetter *et al.* 1995). It is based on establishing 15 models for whole words of different lengths that are trained with all the words in the corpus falling into the same length category. Therefore, they reflect the average acoustic structure of the most frequent words in the corpus.

Several approaches for building generic word models have been reported in the literature, that make use of phoneme models instead of generic acoustic models e.g. (Asadi *et al.* 1990; Asadi *et al.* 1991). Their major advantage is the higher quality in modelling the acoustic properties of unknown words.

Either just arbitrary phoneme sequences of unrestricted length are used or these sequences are restricted in length to a certain maximum. The results reported suggest the conclusion, that it is good to incorporate more information about the acoustic structure of the expected unknown words.

Another advantage of phoneme models over generic acoustic ones is the fact, that the recognized phoneme sequence can be used as an initial phonetic transcription of the detected unknown word.

Enlarging the recognizer's vocabulary

A different approach to detect unknown words is based on enlarging the recognizer's vocabulary (Fetter *et al.* 1995). A certain amount of supplementary words is added to the vocabulary and is said to be one model for unknown words. That is, whenever one of these words is recognized, the hypothesis *"unknown"* is output. By adding many different words a kind of generic word model is built, because of acoustic similarities between the uttered unknown word and some of the supplementary words. The results reported show that this method does not perform better than an approach based on garbage models for whole words.

IMPROVED MODELLING OF UNKNOWN WORDS

The detection techniques presented in the previous section suffer from insufficient modelling quality. This section presents some improved detection methods, that are based on the construction of a generic word model.

Language models and the linguistic context

An analysis of the occurrences of unknown words in a given domain shows, that there are word classes in which unknown words are more likely to appear than in others. Names of people, cities etc. and unknown word forms are more likely than unknown prepositions or names of months etc. So there exists a linguistic context for some categories of unknown words. For that reason it is possible to estimate a statistical language model that takes these categories into account. It is also possible to use this context information as a deterministic constraint (Jusek *et al.* 1994).

Phonotactic restrictions

In each language only a small amount of combinatorial combinations of phonemes are possible to form a valid word. The phonotactic constraints of the language considered describes those phoneme sequences. These constraints can be used to build a more specific generic word model that reflects the acoustic structure of an arbitrary word much better than an unrestricted phoneme sequence (Jusek *et al.* 1994; Jusek *et al.* 1995).

A problem in using phonotactic constraints lies in the fact, that those constraints must be written down by an expert and then be implemented by hand. Therefore, consistent modelling of known and unknown words can never be guaranteed.

Improved acoustic models

All the construction principles for generic word models presented so far use either generic acoustic models or context independent phoneme models. The vocabulary words in a speech recognizer are usually built from context dependent phoneme models e.g. triphones (Lee 1989) or polyphones (Schukat-Talamazzini 1995). These models are highly specific for the acoustic properties of the spoken utterances and therefore reflect the structure of the speech signal very well. Using these models is a significant advantage for the construction of a generic word model. It is, however, very difficult to combine context dependent phoneme models with phonotactic constraints as described in the previous section. A solution to this problem is proposed in the following section.

AUTOMATIC GENERATION OF MODELS FOR UNKNOWN WORDS

In the previous section it was shown that contextual restrictions can significantly improve the modelling and therefore also the detection of unknown words. However, a combination of the last two approaches — phonotactic modelling and context dependent acoustic models — is still very cumbersome if achieved manually. The large training database needed to estimate the parameters of acoustic models for a large vocabulary continuous speech recognizer can, however, be used to automatically derive the most relevant phonotactic restrictions present in the language fragment considered. Usually, for the training corpus an orthographic representation exists. Together with the phonetic transcription of every word from the vocabulary an approximate phonetic transcription of the training set can be created easily. This large string of phonetic units defines how often a specific unit is observed and in which context it occurs. Acoustic models can be estimated for those context dependent units for which enough samples are available. This set of models will generally be a subset of all context dependent units found in the corpus.

The limitation to the use of only a subset of units in representing acoustic events makes a generalisation over the acoustic properties of the training set. However, this generalisation will always be limited i.e. a "context independent" phoneme model will never be truly context independent but restricted to the implicit context of the corpus its parameters were estimated on. Therefore, generic word models based on context independent units are not only inaccurate but also unreliable. Context dependent acoustic units, however, will perform best if used together with the contextual restrictions established by the corpus their parameters were estimated on and will also perform unreliably if combined according to restrictions derived from a different knowledge source by an expert. Therefore, in our approach a model for unknown words is basically generated from the trainable context dependent models combined to a generic word model according to their original contextual restrictions. If the acoustic context can be as large as a whole word and a single sample is enough for training a model's parameters this means that the automatically generated generic word model describes exactly the words of the vocabulary. However, if the contextual restrictions of the models used are generalized also a more general word model can be computed that describes unknown words with similar acoustic properties as the ones found in the training database. Obviously, it will never model words with properties far from those observed but the parameters trainable will also never match these acoustic events properly.

In order to make this method for automatically generating a generic word model more clear let us consider a very simple example. For reasons of simplicity we assume that triphone models are the most specific context dependent models to be used. Let the training material be a sufficient number of realisations of the utterance *"automatic speech recognition"* (/#O+t@+m&+tIk#spitS#re+k@+gnI+Sn#/) so that for all triphones models can be estimated. The sequence is started by the following models:

#/O/+t O+/t/@ t/@/+m . . .

The first one of the above models, the triphone #/O/+t denotes a model of an /O/ with left context word boundary and right acoustic context /t/. The symbol "+" marks a syllable boundary. If exactly these models would be combined to a generic word model according to their contextual restrictions in this case only the words *"automatic"*, *"speech"*, and *"recognition"* themselves would be described as every phoneme pair occurs exactly once leaving no "entry point" for different model sequences. The first abstraction from these context restrictions is achieved by allowing the generic word to start and end at every syllable boundary found. Then also all embedded syllable sequences — e.g. the word *"cognition"* — could in principle be reconized as unknown words.

However, in order to make a true abstraction from the acoustic properties of the training set a further generalisation of the contextual restrictions is necessary. A very simple one is to distinguish between vowels and consonants only. This leads to a set of generalized triphones. The first few models in the sequence are listed below with V standing for vowel and C for consonant:

#/O/+C V+/t/V C/@/+C . . .

A generic word model built from these acoustic models describes many words acoustically similar to the ones found in the training set. In our example among others the following words are described and could possibly be detected reliably as unknown words in an utterance:

speak	/spik/	#/s/C C/p/V C/i/C V/k/#
rich	/ritS/	#/r/V C/I/C V/t/C C/S/#
technique	/teknIk/	#/t/V C/e/C V/k/C C/n/C C/i/C V/k/#
mission	/mISn/	#/m/V C/I/C V/S/C C/n/#

As this example is very restricted and simple many unknown words would be modelled completely inadequately. However, if a generic word model is built from a large training corpus that representatively reflects the acoustic properties of a language a sufficiently general modelling of unknown words can be achieved automatically.

RESULTS

The experiments presented below were carried out on spontaneous German speech in the appointment scheduling domain taken from the VERBMOBIL corpus. We used the ISADORA speech recognition system (Schukat-Talamazzini 1995). The system was trained according to the official 1995 VERBMOBIL evaluation guidelines using about 10.5 hours of spontaneous speech from about 200 speakers covering several different dialects. The recognition vocabulary contained 3304 words plus the generic word model and models for human and non human noises. For the tests we used a bigramm language model, which takes 5 different categories of unknown words into account. The test set consisted of 85 utterances from 42 speakers — about 13 minutes of speech in total. About 9.5% of the uttered words were unknown to the system. These unknown words were mainly proper names, e.g. the names of people and cities.

Table 1 below shows the following measures calculated on the best word chain: ΔWA is the difference of the word accuracy compared to an experiment not using a model for unknown words in the recognition. The *detection accuracy* (DA) was introduced in (Jusek *et al.* 1995) and is defined as follows:

$$DA = \frac{100 * (\text{\# CDUW - \# IDUW})}{\text{\# unknown words in the test set}}$$

Here CDUW denotes the correctly detected unknown words and IDUW denoted the incorrectly detected ones. The detections-correct-rate (DC) is the percentage of correctly detected unknown words, and the false-alarm-rate (FA) is the percentage of incorrectly detected unknown words with respect to all the words in the test set.

The experiments are labelled in the following way: "Phonotactics" stands for a generic word model that was built using phonotactic constraints. "AutoGen" denotes an automatically generated generic word model using polyphone models i.e. acoustic units with context extending to arbitrary length only limited by the training material. For the generic word model the context restrictions were relaxed to use at maximum a single phoneme or phoneme class as left and right context. The additional word "Context" denotes, that the linguistic context of the expected unknown words was deterministically modelled.

Experiment	ΔWA	DA	DC	FA
Phonotactics	+1.6%	10.6%	17.8%	0.7%
Phonotactics Context	+0.8%	12.2%	22.2%	1.0%
AutoGen	+1.2%	0.6%	15.6%	1.5%
AutoGen Context	+3.2%	30.6%	45.0%	1.4%

Table 1: Results of the unknown word detection experiments

The results shown in table 1 demonstrate, that the detection of unknown words can be improved significantly by modelling as much information about the expected words, as possible. The phonotactic constraints are a good way to model unknown words, especially if no linguistic context is available. The automatically generated model achieves much better results if the linguistic context can be restricted. Because it also reflects the acoustic properties of known words extremely well it tends to produce a higher false alarm rate, if linguistic constraints are omitted. In our future research we will focus on reducing the false alarms produced by the automatically generated model.

CONCLUSION

In this paper we presented a new approach to the modelling and detection of unknown words. Its main advantages are that it can make use of the modelling quality achieved by context dependent acoustic units and can also combine this technique automatically with an approximation of the phonotactic restrictions imposed by the language considered. We presented recognition results achieved on a large German spontaneous speech recognition task. These numbers show that the automatically generated model can be used successfully to tackle the problem of unknown word detection in continuous speech.

ACKNOWLEDGEMENTS

This work was funded partially by the German Federal Ministry of Education, Science, Research, and Technology (BMBF) in the framework of the VERBMOBIL project under Grant 01IV102G/7, by the German Research Foundation (DFG) within SFB 360, and by the Commission of the European Communities in the framework of the "Spontaneous Interlingua Network" (SPIN). The responsibility for the contents of this study lies with the authors.

REFERENCES

Asadi, A., Schwartz, R., Makhoul, J. (1990) Automatic Detection of New Words in a Large Vocabulary Continuous Speech Recognition System. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Albuquerque, 125–128.

Asadi, A., Schwartz, R., Makhoul, J. (1991) Automatic Modeling for Adding New Words to a Large-Vocabulary Continuous Speech Recognition System. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Toronto, Canada, 305–308.

Fetter, P., Class, F., Haiber, U., Kaltenmeier, A., Kilian, U., Regel-Britzmann, P. (1995) Detection od Unknown Words in Spontaneous Speech. *Proc. European Conf. on Speech Communication and Technology*, Madrid, 1637–1640.

Hayamizu, S. (1993) Detection of Unknown Words in Large Vocabulary Speech Recognition. *Proc. European Conf. on Speech Communication and Technology*, Berlin, 2113–2116.

Itou, K., Satoru, H., Hozumi, T. (1992) Detection of Unknown Words and Automatic Estimation of their Transkriptions in Continuous Speech Recognition. *International Conference on Spoken Language Processing*, Banff, Canada, 799–802.

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

Jusek, A., Fink, G., Kummert, F., Rautenstrauch, H., Sagerer, G. (1995) Detection of Unknown Words and its Evaluation. *Proc. European Conf. on Speech Communication and Technology*, Madrid, 2107–2110.

Kai, A., Nakagawa, S. (1994) Evaluation of Unknown Word Processing in a Spoken Word Recognition System. *International Conference on Spoken Language Processing*, Yokohama, Japan, 2151–2154.

Lee, K.-F. (1989) *Automatic Speech Recognition: the Development of the S PHINX System*. Kluwer Academic Publishers, Boston.

Sakamoto, H., Matsunaga, S. (1995) Detection of Unknown words using Garbage Cluster Models for Continuous Speech Recognition. *Proc. European Conf. on Speech Communication and Technology*, Madrid, 2103–2106.

Schukat-Talamazzini, E. G. (1995) *Automatische Spracherkennung*. Vieweg, Wiesbaden.