AUTOMATIC TRANSCRIPTION OF UNKNOWN WORDS IN A SPEECH RECOGNITION SYSTEM

R. Haeb-Umbach, P. Beyerlein, E. Thelen

Philips GmbH Forschungslaboratorien Aachen, P.O. Box 1980, 52021 Aachen, Germany
Email: haeb@pfa.philips.de

ABSTRACT

We address the problem of automatically finding an acoustic representation (i.e. a transcription) of unknown words as a sequence of subword units, given a few sample utterances of the unknown words, and an inventory of speaker-independent subword units. The problem arises if a user wants to add his own vocabulary to a speaker-independent recognition system simply by speaking the words a few times.

Two methods are investigated which are both based on a maximum-likelihood formulation of the problem. The experimental results show that both automatic transcription methods provide a good estimate of the acoustic models of unknown words. The recognition error rates obtained with such models in a speaker-independent recognition task are clearly better than those resulting from separate whole-word models. They are comparable with the performance of transcriptions drawn from a dictionary.

1. INTRODUCTION

This paper deals with the problem of adding new words to a speaker-independent recognition system by finding representations of these words as sequences of subword units. In the most elementary case, the spelling of the word is unknown or is not exploited [2]. Thus a phonetic transcription is not available. We assume that we are given sample utterances of the unknown words and an inventory of subword units, which have been trained on a speaker-independent training corpus. These units are to be used to compile the acoustic models of the recognition vocabulary.

The assessment of phonetic baseforms by exploiting acoustic samples may also be used to verify or select phonetic transcriptions obtained from a pronunciation lexicon or a text-to-speech system [3], [1]. Another application is in the context of vocabulary-independent recognition, if only a small-scale application-dependent data collection (few speakers, few tokens) is affordable to complement a large application-independent training corpus. The application-dependent data can be used as 'transcription corpus' for the estimation of sequences of subword units, while the models of the subword units are derived from the large corpus.

We have investigated two methods to automatically transcribe a new word as a sequence of suitably defined subword units. Both methods are based on a maximum-likelihood formulation. In the first method, which is in essence the method described in [2], a separate transcription is first determined for each utterance of a word, and then that transcription is selected that has most likely produced all utterances. In the second method, which we denoted the "average transcription method", an 'average' utterance is obtained from the given utterances of a word. Then this average utterance is mapped onto the subword unit inventory by a standard continuous-speech recognition procedure.

We address the following issues:

- The use of lexical knowledge and acoustical knowledge (dictionary pronunciation or automatic transcription or both).
- The selection of subword units. We compare phonemes with smaller-sized units, so-called context-independent or context-dependent phoneme segments.
- The influence of the number of sample utterances and the lexicon size.

The paper is organized as follows. Section 2 introduces the methods to obtain a sequence of subword units from given sample utterances. In Section 3 the inventories of subword units are described which have been used in the experiments presented in Section 4. Section 5 completes the paper with some conclusions.

2. TRANSCRIPTION TECHNIQUES

It is well-known that the phonetic recognition accuracy of today's automatic speech recognition systems is too low to reliably estimate the phonetic transcription from a given utterance. The estimation can however be improved, if several utterances are employed to estimate

0-7803-2431-5/95 $4.00

a transcription [2], or if the orthography is exploited as an additional knowledge source in order to constrain the search space [3], [1].

Here we concentrate mainly on the first approach, since for our main application the orthography is not available. Two methods have been employed to obtain a transcription (i.e. a sequence of subword units, not necessarily phonemes) of a new word from a few utterances of that word.

Let $y^{(1)}, y^{(2)}, ..., y^{(n)}$ be the n given utterances of the new word w, and let $\mathcal{S}$ be the set of all possible subword unit sequences.

2.1. Multiple-Candidate Transcription Method

In this method, first for each of the n given utterances the most likely sequence of subword units is determined: $T^{(i)}; i = 1, ..., n; T^{(i)} \epsilon \mathcal{S}$, where

$$T^{(i)} = arg \max_{s \epsilon \mathcal{S}} P(y^{(i)}|s). \quad (1)$$

$T^{(i)}$ can be found by a standard continuous-speech subword unit recognition procedure. Then $\hat{T}_{mul}$ is obtained as the one out of the set $\mathcal{T}$ of transcriptions, $\mathcal{T} = \{T^{(i)}; i = 1, ..., n\}$, which has most likely produced all n utterances, i.e. the one for which the product of the likelihoods of all utterances given this transcription is maximum:

$$\hat{T}_{mul} = arg \max_{T^{(i)} \epsilon \mathcal{T}} \prod_{j=1}^{n} P(y^{(j)}|T^{(i)}). \quad (2)$$

This approach is in essence the approach taken in [2] to determine fenonic baseforms.

Note that $\hat{T}_{mul} \epsilon \mathcal{T}$, i.e. only such a transcription can be chosen, which has obtained the best score for at least one utterance.

2.2. Average Transcription Method

In this method, first a separate whole-word model with single-density emission probabilities is trained from the n given utterances. This model can be interpreted as an 'average utterance' $\bar{y}$ obtained from the utterances, where the observation vectors of this average utterance are the mean vectors of the state-specific emission probability densities. The transcription of the unknown word is now given by the subword unit sequence which has most likely produced this average utterance.

$$\hat{T}_{avg} = arg \max_{s \epsilon \mathcal{S}} P(\bar{y}|s). \quad (3)$$

This subword unit sequence can again be found by a continuous-speech recognition procedure.

In contrast to the multiple-candidate transcription method, the selected transcription $\hat{T}_{avg}$ can be any out of the set $\mathcal{S}$ of subword unit sequences.

Note that the average transcription $\hat{T}_{avg}$ can also be used as an additional candidate $T^{(n+1)}$ in the multiple-candidate transcription method.

2.3. Incorporation of Given Phonetic Transcription

Here we are concerned with the problem of how to incorporate one or more given phonetic transcriptions into the approaches presented above. We do not consider the (grapheme-to-phoneme) problem of finding a transcription from a given orthography, but rather assume that the transcription is given, either by a pronunciation dictionary or by a text-to-speech system.

In [1] the search for the phonetic transcription is restricted to the alternatives proposed by a text-to-speech system. Here we take a very straightforward approach: Rather than constraining the search space, the phonetic transcriptions of a dictionary or of a text-to-speech system are taken as additional candidates for the multiple-candidate transcription method; i.e. the set $\mathcal{T}$ is increased by these new transcription candidates. The potential advantage is that unusual pronunciations have a higher probability of being detected compared to the approach of [1].

3. SUBWORD UNITS

The natural choice for a recognition subword unit often is the phoneme since this unit has a direct link with the pronunciation of the word and thus is used in pronunciation dictionaries. For an automatic transcription system, as described here, other units can also be employed.

Our speech recognizer is based on hidden Markov models, the emission probabilities of which are modeled by continuous Laplacian mixture densities with a single 'standard deviation' vector pooled over all states and all mixture components [4]. The phoneme models are 3-'segment' left-to-right hidden Markov models. A segment is a sequence of two states with identical emission probability density functions. The transition probabilities $a(s \mid s')$ for going from state s' to state s are not trained but instead are given fixed a-priori values that are non-zero only for loop, skip and forward transitions. The Viterbi approximation is employed in both training and recognition, i.e. the probability of a word is replaced by the probability of its most likely state sequence.

We have considered the following inventories of subword units:

- 46 context-independent phonemes
- 3*46 context-independent (CI) phoneme segments
- 1274 context-dependent (CD) phoneme segments where a separate model was used for a context seen more than 100 times in the training data. Note that since a phoneme consists of 3 phoneme segments, the context-dependency is either only left-dependent or only right-dependent for the leftmost or rightmost segment of a phoneme, respectively. These units are similar to the phonicles used in [5].

The context-dependent phoneme segments were considered a good compromise between specificity of the models and availability of training material, which in our case was not large enough to model triphones.

Note that transcriptions based on phoneme segments are purely acoustically defined and that they are not based on linguistic or phonetic knowledge.

4. EXPERIMENTAL RESULTS

We trained the subword units on a German telephone speech corpus consisting of 34,000 isolated utterances of 174 speakers. Gender-specific models were trained with 1000 Laplacian densities per gender. We experimented with two different test lexica. The first lexicon consisted of 10 German city names; the second consisted of 32 words of at least 2 phonemes duration, including the 10 city names. The test vocabulary was spoken 3 times by each of 12 male and 12 female speakers, resulting in 720 utterances for the 10-word lexicon and 2304 utterances for the 32-word lexicon. Test lexicon as well as test speakers had no overlap with the training corpus.

Three utterances of each word spoken by 4, 10 or 45 male and the same amount of utterances by female speakers were used to estimate the gender-specific transcriptions. This "transcription corpus" had no overlap with the training and test data.

Table 1 and 2 compare transcription techniques for each of the three subword unit inventories mentioned earlier, for the 10-word and the 32-word lexicon, respectively:

a) Phonetic transcriptions of the 10 or 32 recognition words looked up in a pronunciation dictionary (standard case of vocabulary-independent recognition).

b) Multiple-candidate transcription obtained from the given utterances in the transcription corpus.

c) Average transcription obtained from the given utterances in the transcription corpus.

d) Average transcription used as the (n+1)st candidate transcription of the multiple-candidate transcription method.

e) Transcription found in the pronunciation dictionary taken as yet another, the (n+2)nd candidate transcription of the multiple-candidate transcription method.

Table1: Word error rates [%] of transcription methods as a function of transcription corpus size for 10-word recognition lexicon. Phonemes, CI segments, CD segments and word models (single densities).

Phonemes	# Transcription utterances per word and gender		
	4	10	45
a)	2.9		
b)	1.8	1.7	1.0
c)	1.4	1.0	2.1
d)	1.1	1.4	1.0
e)	3.3	0.8	0.7
CI segments	**# Transcription utterances per word and gender**		
	4	10	45
a)	2.9		
b)	1.8	1.1	0.3
c)	1.1	0.4	0.1
d)	0.8	0.6	0.0
e)	1.7	0.6	0.1
CD segments	**# Transcription utterances per word and gender**		
	4	10	45
a)	2.5		
b)	1.7	1.0	0.8
c)	1.0	1.1	0.7
d)	1.1	1.1	0.7
e)	1.0	1.0	0.8
Word models	12.6	6.9	4.9

The first observation is that speaker-independent word models with single density emission probabilities perform clearly worse than the models obtained from the transcription techniques as a concatenation of subword units. Obviously the transcription corpus is too small to train reliable speaker-independent word models, whereas the concatenated models take advantage of a large speaker-independent training corpus.

Comparing automatic transcriptions with those looked up in a dictionary, it can be seen that the error rates obtained are comparable. In some cases, they even outperform the dictionary transcription. It may be concluded that the automatic transcription methods provide a good estimate of the acoustic models of the unknown words. Smaller subword units tend to provide better transcriptions, in particular the larger the transcription corpus is.

Comparing multiple-candidate and average transcrip-

Table2: Word error rates [%] of transcription methods as a function of transcription corpus size for 32-word recognition lexicon. Phonemes, CI segments, CD segments and word models (single densities).

Phonemes	# Transcription utterances per word and gender		
	4	10	45
a)	3.6		
b)	5.7	5.1	3.5
c)	7.3	4.2	4.7
d)	4.6	4.1	3.5
e)	4.3	3.6	3.4
CI segments	# Transcription utterances per word and gender		
	4	10	45
a)	3.6		
b)	7.5	5.2	4.1
c)	5.2	3.6	2.9
d)	5.2	3.6	2.9
e)	5.0	3.0	2.4
CD segments	# Transcription utterances per word and gender		
	4	10	45
a)	3.6		
b)	5.2	3.6	2.8
c)	1.9	1.8	2.0
d)	1.9	1.9	1.9
e)	2.5	2.3	2.1
Word models	16.5	9.6	5.6

tion there is no unanimous superiority of one of them. A combination of both, however, leads in general to the best performance. It seems, though, that the incorporation of the lexical transcription as additional candidate transcription does not lead to a clear improvement.
A comparison of the results of table 1 and 2 shows that the relative performance of the automatic transcriptions compared to the dictionary transcription is worse for the larger lexicon. This may indicate that the automatically determined transcriptions tend to be less consistent than those of a dictionary.

5. CONCLUSIONS

Two methods have been presented to automatically find a representation of unknown words as a sequence of subword units from given sample utterances of the words.

- The error rates obtained with the acoustically-based automatic transcription methods are comparable to the error rates obtained with dictionary transcriptions and significantly better than those resulting from separate speaker-independent whole-word models with single density emission probabilities.
- If a larger amount of transcription material is available smaller subword units tend to provide better automatically derived transcriptions.

From an implementational point of view it is interesting to note that the developed transcription techniques allow the addition of user-defined words to the recognition vocabulary simply by adding the determined sequence of subword units to the recognition lexicon. In contrast, training separate whole-word models would entail the training and storage of complete new HMMs.

6. REFERENCES

[1] A. Asadi, R. Schwartz, and J. Makhoul. Automatic modelling for adding new words to a large-vocabulary continuous speech recognition system. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 305–308, Toronto, Canada, May 1991.

[2] L. R. Bahl, P. F. Brown, P. V. de Souza, R. L. Mercer, and M. A. Picheny. A method for the construction of acoustic markov models for words. *IEEE Trans. Speech and Audio Processing*, 1(4):443–452, Oct. 1993.

[3] L. R. Bahl, S. Das, P. V. de Souza, R. L. Epstein, R. L. Mercer, B. Merialdo, D. Nahamoo, M. A. Picheny, and J. Powell. Automatic phonetic baseform determination. In *Proc. DARPA Speech and Natural Language Workshop*, pages 179–184, Hidden Valley, PA, June 1990.

[4] H. Ney. Acoustic modelling of phoneme units for continuous speech recognition. In *Proc. 5th European Signal Processing Conf.*, pages 65–72, Barcelona, Spain, Sep. 1990.

[5] L. C. Wood, D. J. B. Pearce, and F. Novello. Improved vocabulary-independent sub-word HMM modelling. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 181–184, Toronto, Canada, May 1991.