CREATING LARGE SUBWORD UNITS FOR SPEECH RECOGNTION

T.Pfau, M.Beham[1], W.Reichl[2], G.Ruske

Institute for Human-Machine-Communication, Technical University of Munich, Arcisstr. 21, D-80290 München, Germany
[1]pc-plus-COMPUTING, Grillparzer Str.10, D-81675 München, Germany
[2]Dialogue Systems Research Department, Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974, USA
Tel.: +49 89 289-28554, Fax: +49 89 289-28535, E-mail: {pfa, rus}@mmk.e-technik.tu-muenchen.de

ABSTRACT

This paper deals with the choice of suitable subword units (SWU) for a HMM based speech recognition system. Using demisyllables (including phonemes) as base units, an inventory of domain-specific larger sized subword units, so-called macro-demisyllables (MDS), is created. A quality measure for the automatic decomposition of all single words into subword units is presented which takes into account the trainability of the chosen units. To create the whole inventory an iterative procedure is applied with respect to the predefined quality measure. Each MDS is represented by a dedicated HMM. By tying the densities of specific phonemes, only the number of mixture coefficients and transitions increases in comparison to the original phoneme models. Recogniton experiments within the German *Verbmobil* evaluation 1996 show that the new simple MDS models are as powerful as standard triphone models, although our MDS models are up to now context-independent.

1. INTRODUCTION

In most state of the art speech recognition systems, the acoustic modeling is performed by HMMs which generally represent subword units. In contrast to the well-known context-dependent phonemes (for example triphones) we want to use larger sized subword units which explicitly contain as many phonemes as possible. The larger the units, the better the coarticulation effects should be handled by the units directly. In many speech recognition tasks training data and application vocabulary belong to the same domain. For these tasks the inventory of SWUs can be chosen according to the words in the training database and according to their frequency. Our aim is to determine this inventory automatically for a given training database by using a predefined quality measure. On the one hand the SWUs must be large enough to capture most of the contextual effects on their realization, on the other hand their frequency of occurrence in the training speech data must be sufficient for a robust estimation of the HMM parameters. It is obvious that larger units, up to whole words or phrases, will be more rare in the training data. So we have to achieve a compromise between the length of the chosen units and their frequency of occurrence.

2. CHOOSING APPROPRIATE BASE UNITS

In contrast to the ISADORA system [1][2], we do not want to select a hierarchical order of several phonetically meaningful layers to represent the phonotactic structure of the subword speech units (SWU). Our units will be created automatically from given base units. By choosing phonemes as base units, both margins of a unit are modeled independently of the phoneme context. This might be a certain disadvantage when compared with triphones which are always modeled context-dependently. To avoid this disadvantage, we start from demisyllables [3] (including all phonemes) as base units to build larger SWUs. Most of the phonotactic constraints are represented implicitly within the demisyllables, which therefore nearly can be modeled context-independently. The inventory of the German demisyllables consists of about 54 initial consonant clusters, about 160 final consonant clusters (the latter can be divided in about 24 „rudiments“ and about 20 „suffixes“), and about 130 vowel clusters including diphthongs and „syllabic consonants“ [3].

The chosen base units are concatenated to particular larger units; we call them macro-demisyllables (MDS). Each of these MDS is represented by a dedicated HMM. The size of such a model varies from a single demisyllable up to whole words or phrases. Our aim is to build large and phonetically highly specific units under the constraint of satisfactory frequency of occurrence in the training data to achieve statistical significance during HMM parameter estimation. For speech recognition systems like the German *Verbmobil*, for which the training speech data and the application vocabulary are from the same domain, it is possible to automatically build an inventory of MDS, which is in a certain way optimal with respect to both specific modeling and trainability.

3. CONSTRUCTION OF A MDS INVENTORY

It is not possible to create all sets of MDS which cover the given vocabulary and to choose the one set which

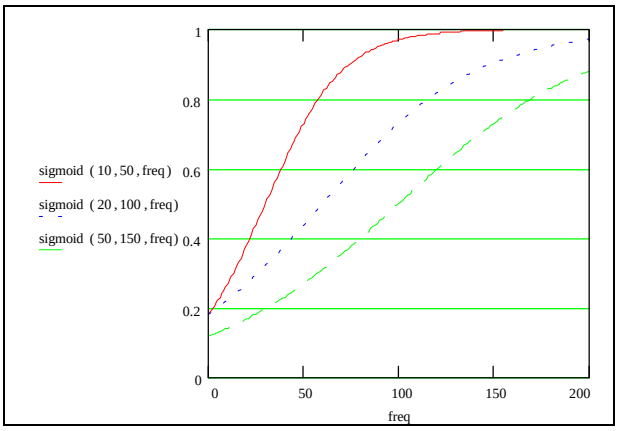

Figure 1: Sigmoid function for the determination of the quality *q* of a single MDS

meets a certain global optimization criterion. Therefore we developed a suboptimal iterative procedure to decompose the given vocabulary in a set of MDS that maximizes a predefined quality measure. The quality measure for a single word has to take into account both the frequency of occurrence of all those MDS, which are used to represent the whole word and in addition the number of MDS used (see Figure1).

We define a value

$q(MDS_k) = sigmoid(F_{min}, F_{max}, freq(MDS_k))$

with $0 < q < 1.0$ as the rating of the frequency *freq* of the unit MDS_k in the training data, which rises strong monotonously with its frequency.

The meaning of the two specified parameters F_{min} and F_{max}, which fix the turning point of the sigmoid function, is the following: a particular MDS is entered in the inventory if and only if its number of occurrence in the training data exceeds a lower bound F_{min} in order to guarantee statistical stability during HMM training. An upper bound F_{max} leads to a saturation of *q*, making sure that all MDS exceeding this limit have nearly the same value. The quality measure for one specific decomposition of the whole word W_i in *k* MDS is simply defined as the product

$$Q(W_i) = \prod_{k \in W_i} q(MDS_k)$$

The qualities *q* of all *k* MDS build the quality of the whole word. The average of the qualities of the whole vocabulary can not be maximized exactly. This is the reason why we developed an iterative algorithm which maximizes the quality *Q* of word W_i independently of all other words. Roughly, the construction of a MDS inventory proceeds as follows:

1. Build all possible MDS given in the training data with sizes from demisyllables alone (including all single phonemes) up to whole words.

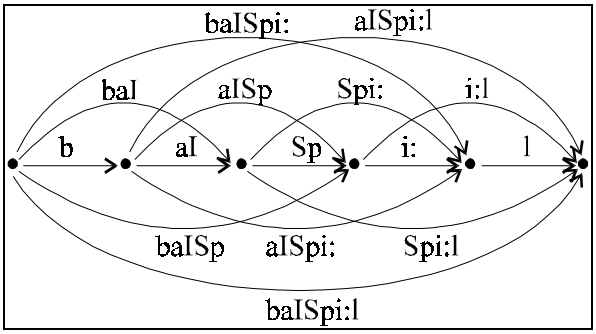

Figure 2: All possible decompositions of the word „Beispiel“ into demisyllables

2. Calculate the frequency rating of every MDS by using the number of occurrence in the training speech data.
3. Eliminate those MDS for which the frequency of occurrence does not exceed the lower bound *Fmin*.
4. Determine the decomposition of every word by maximizing the product $Q(W_i)$. During this step all frequency ratings $q(MDS_k)$ of the MDS remain constant regardless of the use of these MDS in other words. This maximization can be performed by a recursive algorithm similar to the Viterbi segmentation. Figure 2 shows all possible decompositions of the word „Beispiel“ into demisyllables. Every path through the graph, displayed in Figure 2, represents one possible decomposition of the word. During the maximization process the path with the highest quality measure $Q(W_i)$ is searched using the Viterbi algorithm.
5. Count only the MDS actually used after the decomposition of the vocabulary and calculate a new $q(MDS_k)$ for all these macro-units.
6. Proceed with step 3.

In the tested *Verbmobil* database (CDs1-5, CD7 and CD12) about 40000 possible MDS of sizes from one up to 15 phonemes can be found. About 5000 units meet the lower frequency bound criterion of step 3. With the chosen parameters (see Table 1) a stable inventory of 595 MDS is determined after 7 iterations. This can be seen as an optimal set to decompose the 7000 word training vocabulary with respect to the predefined quality measure.

F_{min}	F_{max}	number of iterations	number of units
20	50	8	738
20	100	6	607
50	100	7	595
100	500	4	264

Table 1: Number of iterations until a stable set of units is reached

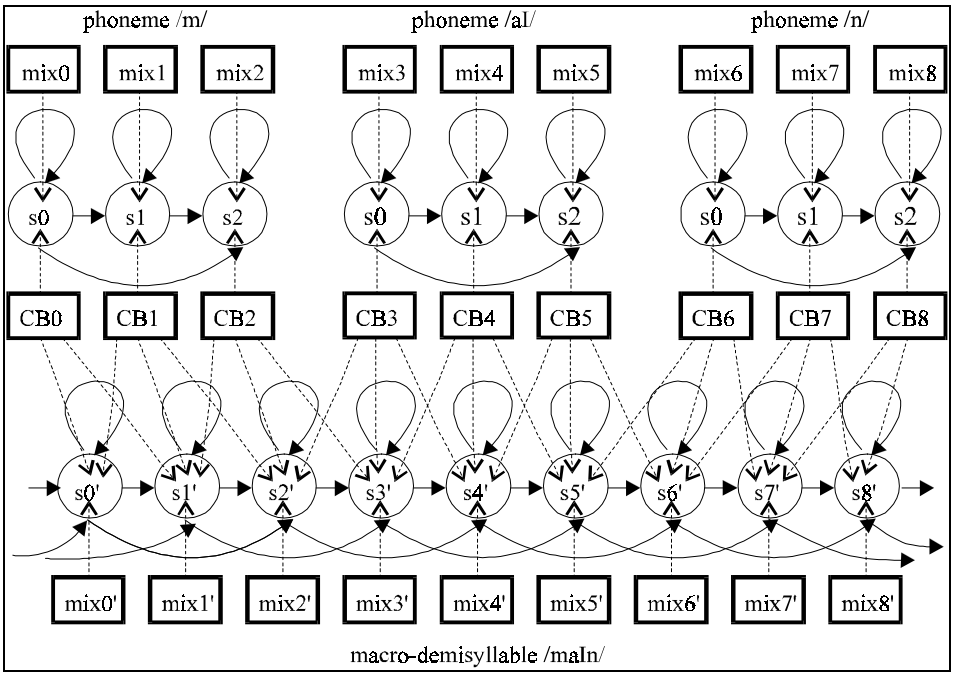

Figure 3: Concatenation of the phonemes /m/, /aI/ and /n/ to the MDS /maIn/ using the densities (codebooks CB0-CB8) of adjacent states (radius 1)

4. SETUP OF MDS HMMs

The macro-unit HMMs can be initialized by simply concatenating the states of all corresponding phoneme HMMs. This leads to a dramatical rise in the number of states in the inventory from about 170 (phoneme HMM) to about 7300 states (MDS-HMM). Using continuous densities (gaussian mixtures), we limit the number of total densities to those of the original phoneme HMM models by tying the densities of the specific phonemes. To allow more flexibility in comparison to the simple concatenation of the states with their densities, the states of the resulting MDS-HMMs can not only use their own densities included in their own phoneme-specific codebooks but also the densities of adjacent states within a predefined radius of influence.

Figure 3 shows an example for the MDS of the German word „mein“ (/m aI n/). The basic phoneme models (/m/, /aI/ and /n/) consist of three states (s0, s1, s2) each. These states are combined to a nine-state sequence (s0'...s8') representing the MDS „mein“. In the upper part of Figure 3 the original phoneme HMMs with continuous densities (codebooks CB0...CB8) and specific sets of mixture coefficients (mix0...mix8) can be found. The resulting MDS-model is represented in the lower part. The number of states simply results from the sum of all phoneme states. In addition to the original codebooks - used in the corresponding phoneme states - the new states have access to the codebooks of the neighboring states, too. The use of the neighboring codebooks depends on the defined radius of influence (radius 0 for example represents the original condition). In this way the radius controls the „focus“ of the individual states. As a consequence, the mixture sets (mix0'...mix8') of the new states have to be enlarged corresponding to the number of densities in the neighboring codebooks. Figure 3 shows the conditions when using radius 1. For example the mixture set mix1' in state s1' refers to the codebooks CB0, CB1 and CB2. Taking into account the neighboring codebooks the HMMs are able to deal with coarticulation effects directly.

Together with the transition probabilities the MDS units thus allow to model the individual lengthening or shortening (up to an omission) of phonemes in the context of the adjacent phonemes, when the radius of influence is chosen appropriately. The mixtures are set during the training procedure according to the individual conditions within each MDS-model. Pursuing our simple tying strategy only the number of mixture coefficients rises when compared to the phoneme models. The number of mixture coefficients used rises from about 20000 (phoneme HMMs) to about 1.14 Mio (MDS radius 1) after some training iterations. After initialization the MDS-HMM parameters are reestimated using Viterbi training, including the removement of very small mixture coefficients. This reduces the number of parameters in the resulting MDS-models significantly.

5. EXPERIMENTAL RESULTS

In this paragraph some of our recognition experiments made with our macro-demisyllable models are presented. To evaluate the performance of our MDS-models the MDS-model approach participated at the

evaluation 1996 [4] of the German *Verbmobil* task. The training database of the evaluation 1996 task consists of about 27 hours of spontaneous speech dialogues in the field of timetable appointments, and the test database consists of about 41 min of spontaneous speech dialogues in the same domain.

The training is carried out using the standard Viterbi algorithm. The search engine uses a tree-based lexicon with one single pronunciation per word and a beam search algorithm including a histogram pruning procedure for means of efficiency.

The experiments were carried out with our new MDS models and - for comparison - with standard triphones as well as with the original context-independent phoneme models (see Table 2).

Basic phoneme HMMs for building the MDS-models:
54 context-independent continuous phoneme HMMs (three or four states each) with a total number of 176 states and 161 codebooks (some states share the same codebooks). The codebooks use a total number of 18808 diagonal gaussian densities of 66 dimensions. The HMM transitions between the states of the HMMs are trained individually for each model-state. The total number of mixture coefficients is about 20000.

MDS models:
595 continuous MDS-HMMs (from three up to 39 states) with a total number of 7273 states using the newly estimated densities of the original 161 phoneme codebooks with 18808 diagonal gaussian densities of 66 dimensions.

Triphone models:
2100 triphone models (inclusive across word modeling) chosen by their frequency of occurrence.

type of models	word rec. rate	word error rate	total no of parameters
phonemes	69.2%	35.4%	2.5 Mio
MDS (radius 1)	78.2%	25.8%	3.6 Mio
triphones	78.5%	25.2%	3.5 Mio

Table 2: Recogniton results on the Verbmobil evaluation 1996

These first results show that simple context-independent macro-demisyllables are almost as good as context-dependent triphones. The efficiency of MDS units can be further increased, when additionally the context at the unit boundaries is taken into consideration. This is already done by triphones. Thus the MDS units seem to be a potential basis for further improvement. These experiments are currently under investigation.

6. CONCLUSION

In this paper a method for selecting domain-specific larger sized subword units for speech recognition systems is presented. It is an automatic, iterative and data-driven method for the selection of appropriate context-freezing units by using a quality measure based on the frequency of occurrence of the chosen units. The increase in the number of parameters is limited - by a simple tying mechanism - to an increase in the number of mixture coefficients. By this means the number of parameters increases by 44% and the word error rate is reduced by 27% both relative to the phoneme models.

Our future work in this area will concentrate on the modeling of the context-dependency of the chosen MDS-units. First the left and right margins of each MDS-unit can be modeled context-dependently similar to the common triphone or polyphone approach. Another aspect is that the codebooks, which are used in the MDS-states, are still phoneme-specific. This disadvantage can be overcome by introducing MDS-specific codebooks. Then appropriate tying mechanisms must be applied to reduce the increasing number of parameters.

Another possible area for improvement is the mechanism of decomposition of words into SWUs. Not only the frequency of occurrence of a subword unit but also the likelihood (maximum likelihood strategy) in the training process should be considered by the quality measure during the iterative selection procedure. Thus the set of units could be selected with respect to the best increase in the overall likelihood.

ACKNOWLEDGEMENTS

This work was funded by the German Federal Ministry for Research and Technology (BMBF) in the framework of the *Verbmobil Project.*

REFERENCES

[1] E.G.Schukat-Talamazzini, H.Niemann, W.Eckert, T.Kuhn, S.Rieck; *Acoustic modelling of subword units in the ISADORA speech recognizer*, in Proc. ICASSP'92, San Francisco, 1992, Vol. 1, pp 577-580.

[2] E.G.Schukat-Talamazzini, *Automatische Spracherkennung, statistische Verfahren der Musteranalyse*, pp 271-316, Vieweg-Verlag, 1995.

[3] B.Plannerer, G.Ruske, *A continuous speech recognition system using phonotactic constraints*, in Proc. of EUROSPEECH'93, Berlin, 1993, Vol. 2, pp 859-862.

[4] J.Reinecke, *Evaluierung der signalnahen Spracherkennung im Verbundprojekt Verbmobil (Herbst 1996)*, Memo 113, Verbmobil, DFKI Saarbrücken, 1996.