SPEECH RECOGNITION FOR SPONTANEOUSLY SPOKEN GERMAN DIALOGUES

Kai Hübener, Uwe Jost, Henrik Heine

Kai Hübener
Siemens AG, ZT SN 5
Kai.Huebener@zfe.siemens.de
Uwe Jost, Henrik Heine
Hamburg University, FB Informatik
{jost,heine}@informatik.uni-hamburg.de

ABSTRACT

This paper presents a HMM speech recognition system for spontaneously spoken dialogues. It has been developed as part of the German Verbmobil project whose aim is the development of a translation support system for face–to–face conversations. The HTK-based decoder deals successfully with some of the hard problems of recognizing fluent speech and reaches quite competitive recognition results.

1. INTRODUCTION

One of today's most challenging areas in automatic speech recognition is the recognition of spontaneously spoken dialogues[1]. Key problems in this field that have to be solved include handling of false starts, hesitations, pronunciation variants, multi–sentence utterances with long silence intervals, non–grammatical constructs and unknown words. This paper presents a speech recognition system, developed in the framework of the German Verbmobil project, which addresses some of the above issues. Verbmobil is a long–term project on the translation of spontaneous language in negotiation dialogues [3].

First, the overall system is sketched. Some specifics of the system are explained in more detail and recognition results are presented. The paper concludes with a short analysis of the most frequent recognition errors made by the system.

2. THE RECOGNITION SYSTEM

Our recognition system is based on a commercially available HMM Toolkit developed at Cambridge University [4].

2.1. Preprocessing

The speech signal is transformed to a sequence of 39–dimensional feature vectors consisting of 12 mel–spaced cepstral coefficients plus normalised log energy along with their first and second order differentials. The frame rate is 100 Hz. Since speaker boundaries are unknown for the test sets, supervised speaker–adaptation techniques cannot easily be applied. We use an utterance–based speaker–normalisation scheme instead. This technique, called state–based cepstral mean subtraction (SCMS), is based on vector quantisation using a codebook consisting of Gaussian means $\mu_1, \ldots, \mu_N$ derived from some initial HMMs. In all experiments the codebook consisted of 165 codewords.

$$\mu^* = \text{argmin}_{i=1,N} \|\mathbf{o}_t - \mu_i\| \quad (1)$$

$$\hat{\mathbf{d}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{o}_t - \mu^* \quad (2)$$

$$\hat{\mathbf{o}}_t = \mathbf{o}_t - \hat{\mathbf{d}} \quad (3)$$

Each feature vector $\mathbf{o_1}, \ldots, \mathbf{o_T}$ of an utterance is then quantised (1) and the overall quantisation distortion for the utterance is computed by adding the squared Euclidian distances to the nearest codebook centroid for each dimension. This distortion vector is then averaged by dividing it by the number of frames (2) and subtracted from each feature vector (3). The normalised feature vectors $\hat{\mathbf{o}}_t$ are then subjected to an LDA transform [2] and reduced to 28 dimensions.

2.2. HMM Topology

Except for the noise models, all HMMs have three emitting states and a strict left–to–right topology. Transition probabilities are also reestimated during model training. There are five different silence models, i.e. utterance–initial, within–utterance and utterance–final silence and between–word short pauses and filled pauses. The first three models have three emitting states each, but are fully ergodic. The rationale for this is that silence does not have a temporal structure as opposed to phones. The between–word silence model is a single–state model which may optionally be skipped because short between–word pauses are rather seldom in fluent speech. The model for filled pauses has five emitting states since filled pauses tend to be much longer than the average phone.

2.3. HMM Training

HMM training proceeds in a number of stages. First, initial monophone models are estimated using a phonemic segmentation of a subset of the the complete training data to

an orthographic transcription. In this step a pronunciation dictionary with up to 300 variants per word is used.[1] The resulting phone labels are then used for training a new set of monophones. This set is cloned to train single–mixture cross–word triphones. Also included are word–dependent triphones for frequent function words. The triphone states are tree-clustered using phonetical questions concerning the triphone context.[2] This reduces the number of distinct HMM states by almost 94% to about 2.300, from which 5.200 (physical) triphones are formed. The other 7.000 (logical) triphones seen in the training process are mapped to one of the physical ones. After clustering, the number of Gaussians per state is iteratively incremented (up to 8) and retrained.

Our silence models are context–independent and much more parameters can be trained for them than one could train for triphone models. Therefore the number of Gaussians for the silence models is twice the number of densities of triphone models. For recognition, all unseen triphones are created using the decision trees produced in the clustering process.

2.4. Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a standard technique known from the field of pattern classification (see [2]). It allows to reduce the number of features needed to distinguish patterns by computing a suitable transformation of the feature space.

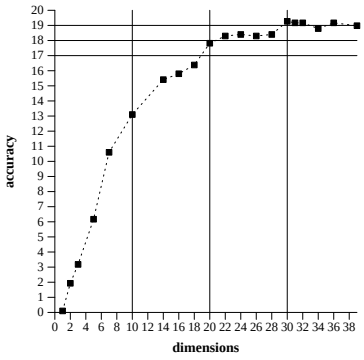

Figure 1: word accuracy per number of LDA dimensions

Figure 1 shows the word accuracy as a function of the number of dimensions retained after performing the LDA matrix multiplication. Due to limited resources, it was impossible to train and test a full system for each dimension. Instead, only a single–mixture monophone system was trained using the full LDA transformation matrix. Recognition experiments were performed on a small test corpus after reducing the number of dimensions of the mixture densities of the trained HMM set. It can be observed, that the number of dimensions can be reduced by almost 50% when a small loss in the accuracy is acceptable. To use more than 30 dimensions does not seem to further improve the recognition rate. Our triphone system uses 28 target dimensions and reaches about the same recognition rate as a comparable system without an LDA in our experiments.

According to the forced alignment of our test corpus, the average number of frames per triphone state is just 2.5 (25 msec), due to the high speech rate of fluent speech and certain performance phenomena like elisions and word contractions. This suggests that the frame context is modelled appropriately by the delta and acceleration coefficients and no further improvement should be expected from using super–vectors of neighbouring frames. Preliminary experiments with supervectors of context size two and three support this conclusion.

2.5. Skip Models

In spontaneous German speech, the duration of a phone is usually much shorter than in read speech. Thus our corpus contains a significant proportion of phones that were shorter than three frames (30 ms) — the minimal duration of a phone according to the HMM topology as described in 2.2. Even some words were shorter than three frames (e. g. *ich* (I)) and 2.5% of all words were shorter than the minimal number of frames according to HMM topology and dictionary. This is not only due to the high speech rate but also to certain performance phenomena, e.g. @–elision.[3]

To model this phenomenon, two skip transitions with a small probability were added to all non–silence HMMs, thus allowing the states 1 and 3 to be skipped. This led to a significant increase of the computational cost of the recognition process due to the higher number of possible alignments.

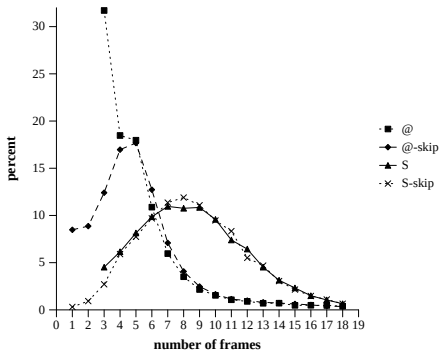

Figure 2: Frame duration for `[@]` and `[S]` from forced alignment using standard topology and skip–models.

Figure 2 shows the duration distribution for the phones `[@]`

[1] We would like to thank the University of Bielefeld for generating and supplying this dictionary.

[2] We wish to thank the IKP at the University of Bonn for their help with designing the questions.

[3] A typical phenomenon in spoken German, e.g. the words *Tagung* (meeting) and *Tagungen* (meetings) will both be pronounced `[t a: g U N]` even though the standard pronunciation for *Tagungen* is `[t a: g U N @ n]`.

and [S] and their skip–counterparts taken from the forced alignment of the German spontaneous speech training corpus (100Hz frame rate). It can be seen that the significance of the problem of short durations is clearly dependent on the phone. Closer examinations revealed that using skip models for a rather small number of triphones could account for almost all cases of short phones.

3. RESULTS

3.1. Verbmobil Evaluation

Each year, all decoders developed in the framework of the Verbmobil project are evaluated. Since the conditions of these evaluations are well defined and the recognition results are available from different sites, we use the setting of the 1995 competition not only to demonstrate the performance of our baseline system, but also to meassure the improvement of our current system compared with the previous version. In last years competition, two evaluation sets were distinguished. One (EVAL95s) contained 265 short sentences (less than 15 seconds each), the other one (EVAL95l) comprised 66 long sentences. Each set contained about 23 minutes of speech. The out–of–vocabulary rate was fixed at 3%. Both test and training data was collected at four sites using a standard high–quality microphone. There were some 200 dialogues totalling to 10 hours of speech in the training set.

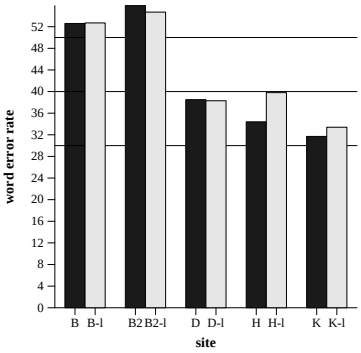

Figure 3: word error rates in 1995 Verbmobil evaluation

Figure 3 shows the word error rates for both evaluation sets. The dark bars represent the error rates for the short test set, "H" stands for "Hamburg University" and "K" for "University of Karlsruhe".[4] It can be seen, that the error rates of the 1995 version of our decoder were close to the lowest error rates reached in the evaluation for the short test set while the results for the long set were about average.

The performance of the current version of the decoder is compared with the perfomance of the older version in figure 4. The error rate could be reduced significantly for both categories. The remarkable difference in the improvement between the category "short" and "long" can be explained by the improved silence modelling in the recognition network. In the older version, a longer silence often triggered the "final silence" model and words following this silence were deleted. In the current version, this is avoided by mapping the "final silence" to the same symbol in the word net as the normal silence model.

[4] For more details see http://sbvsrv.ifn.ing.tu-bs.de/eval.html and http://werner.ira.uka.de/~monika/pages/VM-Eval-1995.html.

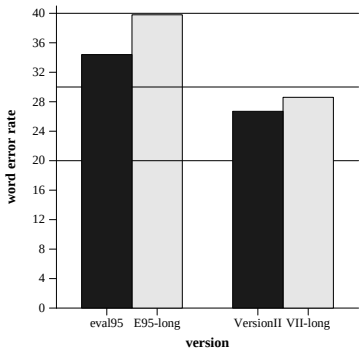

Figure 4: word error rates 1995 vs. 1996

We estimate that using tree–clustered cross–word triphones accounts for about 2–3 percent points of the improvement, with the careful selection of the questions used for the clustering process playing a very important role. The state–based cepstral mean subtraction can reduce the word error rate by about one percent. Using models with skip transitions for selected triphones contributed just over one percent to the gain in the recognition rate, while the usage of more elaborated silence models contributed less than one percent. Using the LDA decreases the computational cost but not the error rate.

4. DISCUSSION

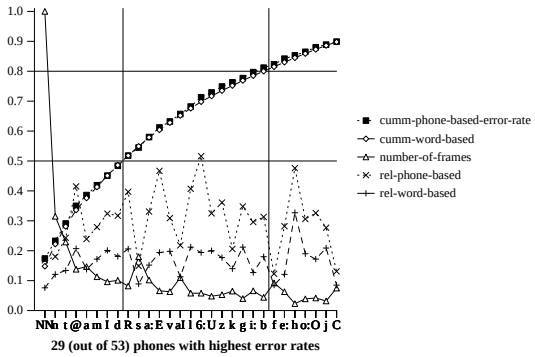

Figure 5: phone recognition errors

Figure 5 shows the relative frequency of phone recognition errors for each phone on a frame–by–frame basis for the 29 phones with the highest error rates. The phone transcription was produced with skip triphone models using forced alignment. Recognition experiments were performed both

word–based (keeping track of the phone models) and purely phone–based. The "cumulated" error rate graphs represent the total proportion of the errors on the overall error, while the "rel" (relative) errors were calculated as proportion $\frac{\#frames\ not\ recognised}{\#frames}$ per phone. The values in all categories were divided by its respective maximal value.[5] Hence the "1.0" on the y-axis represents 72.500 in the case of the "number of frames" labelled with the respective phone, 37.400 for the cumulated word–based and 70.000 for the cumulated phone–based error rate, and 100% for both the word and phone–based relative error rates. The overall number of frames was 280.000.

It is obvious, that seven phones plus the non–speech models `[NN]` (together roughly 16% of all models) are responsible for 50% of all recognition errors.

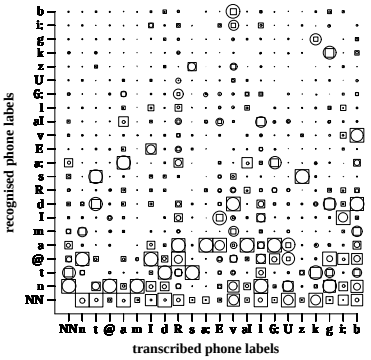

Figure 6: phone confusion matrix

Figure 6 shows the phone confusion matrices for 22 phones (that account for 80% of all errors) for word and phone recognition. The diameters of the circles represent the number of frames that were wrongly labelled with the respective phone on the y–axis when a particular phone (shown at the x–axis) was transcribed. The scaling factor was chosen so as to give the largest circle in each column the size 1. Similarly, the side length of the squares stands for the frame error rate for the phones as recognised during word recognition. It is remarkable, that almost all phones in the matrix are mistaken to belong to a non–speech category rather frequently in the case of word recognition (large squares in the lowest row) while this case seems to be rather rare when phone recognition is run. A possible explaination is, that the language model (in connection with a high grammar scale and word penality) favors long silences.

The mutual confusability of the phones `[n @ m]` is certainly due to the effects of @-elisions. That a `[t]` is often mistaken for a `[d]` is probably the consequence of terminal devoicing, that is modelled in the lexicon but not always observable in fluent speech due to other effects (e. g. assimilation). Our lexicon contains many words where sequences of `[t]` and `[s]` are transcribed (e. g. *Arzt* (doctor) `[a:6 t s t]`). It will hardly occur in fluent speech that each of these phones is really pronounced. Therefore `[t]` and `[s]` are often mistaken.

It seems, that the errors with the major impact on the recognition rates can hardly be avoided by more detailed triphone modelling or more sophisticated signal processing methods. To improve the recognition rate significantly, it would be necessary to model the performance phenomena that can not be handled by triphone models.

4.1. Compound Words

In the current version of the decoder, a full–form pronouncation lexicon is used. This is no problem as long as a limited vocabulary is used. For a practical application, however, it would not be acceptable to allow, for instance, *Junitermin* (appointment in June) but not *Julitermin* (a. in July) in the vocabulary. Trying to account for all possible compound words that can be formed from a basic set of German words by just adding entries to the dictionary would be impracticable. Furthermore, in fluent speech, word compounds are often interrupted by hesitations or non–speech noise.

At the University of Bielefeld, a morphological module is being developed that allows to find potential compounds in word graphs, even if they are interrupted (see [5]). We performed preliminary experiments using an early version of the module and a reduced test set with a relatively small vocabulary. First results suggest, that the reduction in the computational cost of the decoding is significant while the sentence recognition rate of the word graphs can be improved.

5. REFERENCES

1. Ron Cole et al. The challenge of spoken language systems: Research directions for the nineties. *IEEE Trans. Speech and Audio Processing*, 3(1):1–20, January 1995.
2. Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Academic Press, San Diego (CA), 2nd edition, 1990.
3. Wolfgang Wahlster. Verbmobil — translation of face–to–face dialog. Research Report RR–93–34, DFKI, June 1993.
4. Steve J. Young. The HTK Hidden Markov Model toolkit: Design and philosophy. Technical Report CUED/F–INFENG/TR–152, Cambridge University Engineering Department, Cambridge (UK), September 1994.
5. Frederek Althoff, Guido Drexel, Dafydd Gibbon, Harald Lüngen, Martina Pampel, Christoph Schillo. 1996. "Morphology and Speech Technology." To appear in: *Proceedings of ACL 1996*, University of California, Santa Cruz.

[5] The maximal relative recognition errors were reached by very infrequent phones not shown in the figure.