WHAT IS WRONG WITH THE LEXICON – AN ATTEMPT TO MODEL PRONUNCIATIONS PROBABILISTICALLY

Uwe Jost, Henrik Heine and Gunnar Evermann

University of Hamburg
Vogt-Kölln-Str. 30
D-22527 Hamburg
{jost|heine|3everman}@informatik.uni-hamburg.de

ABSTRACT

We motivate the integration of a probabilistic pronunciation model into a system for recognizing spontaneous speech and propose a possible architecture of such a model. In order to develop an environment for experiments, a simplified version employing constrained phone recognition and discrete syllable-size HMM subword units was implemented and evaluated. Although the results are still significantly worse than those achieved by our "conventional" word recognizer, they are encouraging given that the experimental system is only a coarse approximation of the proposed approach.

1. MOTIVATION

In spontaneous speech, a significant portion (more than 40% according to our data) of words (tokens) is not pronounced the way a standard dictionary would predict. This is mainly due to various forms of coarticulation, but also caused by the accent or personal preferences of the speaker. Of course, triphone HMMs can successfully handle a significant amount of variability, provided the lexica used for training and recognition are the same. On the other hand, robust, i.e. noisy models will not discriminate as well as more specialized ones. We even observed that our phone recognizer (trained using a canonical pronunciation lexicon) frequently detected phones that were not articulated but somehow "suggested" by the context. Furthermore, while there will often be a strong regularity of the modifications the context causes to a phone, each phone is also a context for some other phones and may cause confusion for them (e.g. the @-elision in the /t- +n/ context will cause the model /t-@+n/ to be trained on "not being there" but the following /@-n+*/ models cannot be well trained since the actual acoustic context is not /@-/ but /t-/). Even with rather noisy triphone models, we found a significant difference between the recognition rates for canonically pronounced words and alternatively pronounced ones. On a corpus of 175 (manually phone-labeled) utterances, the word recognition rate (correctness) of the 87 utterances with a higher proportion of canonically pronounced words (more than 73%) was 86%, while the recognition rate for the other sentences was only 80.5%.

These considerations suggest that an improved lexical model could potentially lead to significantly higher recognition rates because it would allow to recognize a much larger portion of non-canonically pronounced words and to train better acoustic models.

Adding frequency-weighted pronunciation variants to the lexicon is certainly a step in the right direction but for larger vocabularies there are serious problems concerning the training of such variants and the size of the resulting search space.

An ideal lexical model would attach a probability to any pronunciation of each word (or even phrase) and be trainable using a limited amount of **word**-transcribed data.

2. A PROBABILISTIC LEXICON

The most obvious choice for such a model (for someone working with HMM speech recognizers) would be an HMM itself, resulting in a word recognizer consisting of a two layered HMM model; the first (acoustical) layer containing the phone models and the second (lexical) layer modelling pronunciations of words (see Figure 1). The

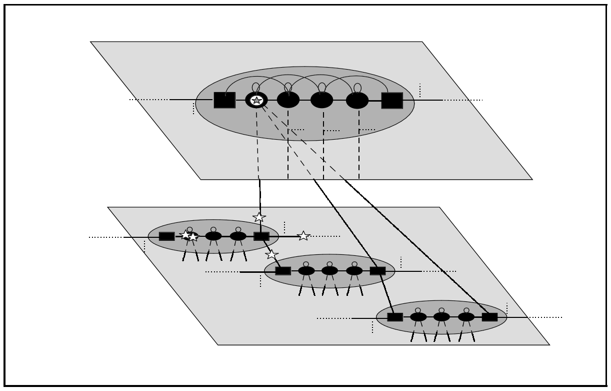


Figure 1: Two–layer model

This research was partly funded by the Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the VERBMOBIL Project under Grant 01 IV 101 A/O.

working of such a model can best be explained using the token–passing paradigm [10]. The first layer consists of a triphone (or n–phone) HMM network as commonly used for phone recognition. Phone sequence hypotheses (paths) are represented by tokens that are propagated through a network of HMM states. Each state must be able to hold multiple tokens to avoid search errors.

Whenever a token is propagated to the final (non–emitting) state of a phone HMM, the token is duplicated and passed to the entry states of all connected phone HMMs in the usual manner, adding phone language model scores and phone transition penalties to the score of the token. Additionally, the token is also passed to all connected HMM states in the upper level – the lexical layer. This layer, however, uses its own token set. Hence the token passed from the lower level is not treated as a token anymore but as an output symbol. Provided there is a token in the corresponding state in the higher level HMM and a certain time constraint is met, the emission probability of the symbol is added to the score of the token, or put more precisely: to the proportion of the token's score that was accumulated in the last acoustical HMM. Then the token is propagated in the usual way.

In theory, every state in every higher level HMM would have to be connected to the final states of all acoustical HMMs, to ensure that the probability of every pronunciation of every word can be calculated. This should not pose a real problem, however, since the network can be reduced heavily in various ways. First, it would not make sense to distinguish between all context dependent versions of an allophone when a token is passed to the upper layer, i.e. almost all connections can be bundled for each allophone. Furthermore, HMM states in the upper level should be shared to a large extent (possibly using tree–based clustering methods). Finally, the network can be pruned.

The pronunciation models in the upper layer could be based on word or subword units. We examined the number of different words, morphs and syllables being observed when using an increasing amount of our Verbmobil 96 training data. Figure 2 shows that the increase in the number of words is still linear even after having observed 200,000 words. This translates to an OOV rate of about 1.1% on the last 50K words of the training set. A linear increase is also observed for the number of morphs even though all training data was from one domain only. Since the syllabic subword units yield the smallest inventory among these alternatives, we decided to use them to model pronunciation phenomena. Given that the number of potential syllables in German is finite (although rather large) one can hope to be able to train almost all syllables that are needed for recognition and thus have sensible pronunciation models even for words which were not in the training set.

Independent of the choice of subword units the proposed two-layered model could be quite expensive computationally. This could make it very difficult to conduct the experiments necessary to work out its details, to understand its working and find its weaknesses.

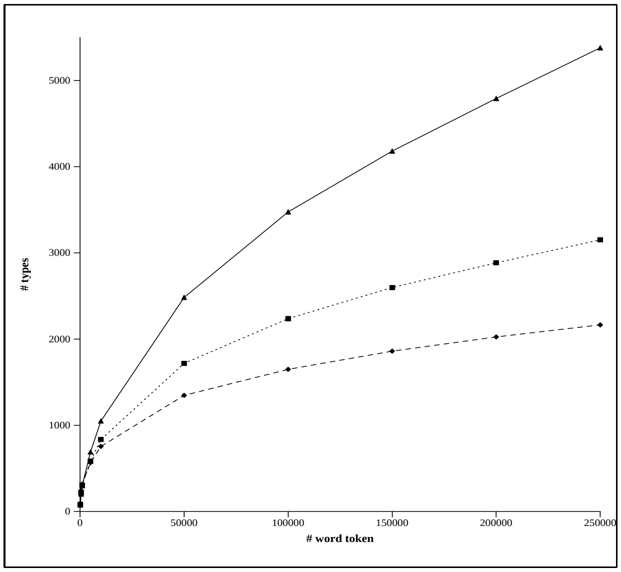


Figure 2: The number of different words (upper graph), morphs (middle) and syllables (lower) for increasing amount of training data being used.

3. SIMPLIFIED MODEL

In order to get a simple model to experiment with, we decided to first implement a two-step HMM model where the second layer is not directly connected to the first one but only the (1-best) output of the first layer forms the input to the second one. Since the output of the first layer is a string of phones, this layer can quite easily be trained and evaluated separately.

With this experimental setting, we mainly want to answer the following questions:

- What should the architecture of the lexical models look like?
- Which words and phrases should get their own model and how can they best be found automatically?

The two-step simplification comes at a cost – we can not expect this model to find the same globally optimal solution that the integrated two-layered model could find. We investigated the loss of information caused by only propagating the best phone string to the second stage.

The two-layer model can be simulated by generating a huge phone-lattice and performing the second step on this lattice. Based on these lattices we estimated a-posteriori probabilities for each phone on a frame by frame basis. We measured the amount of information necessary to predict the probabilities of all phones given the identity of the best phone. The Kullback-Leibler distance was used to measure the distance of the average distribution given the best phone and the actual distribution.

The (weighted) average of these distances is about as high as the entropy of the average distributions. This means that we are losing about half the information contained in the lattices.

To get a feeling for the upper bound on the recognition rate of a two-step recognizer, we replaced both recognizers by humans, i.e. we used manually phone-labeled utterances and presented these phone sequences to a test-person who tried to “recognize” the word sequences. The achieved word accuracy of about 94% is probably close to the level of agreement two human transcribers would reach on the same corpus. This suggests, that a “perfect” phone recognizer could serve as a suitable preprocessing module – its output would still contain enough acoustical evidence to recognize the utterance. The 20-30% error our HMM phone recognizer currently introduces, however, translates to an accuracy of only 73% for the human “word recognizer”. These recognition rates are only upper bounds because the persons who phone-labeled the data probably used lexical and semantic knowledge, which is not available to our first stage. The second stage “human recognition” also employed semantic knowledge, which would be very hard to encode in our HMM models.

4. FIRST IMPLEMENTATION & RESULTS

4.1. Acoustical layer

Our baseline HMM recognizer was developed in the framework of the Verbmobil project, aimed at developing a speech–to–speech translation system for German dialogues. The decoder was tailored for word recognition and features tree-clustered cross-word triphones with 14–mixture Gaussians [2]. In last years Verbmobil acoustic evaluation, the recognizer reached a word accuracy of 80% (83.2% correctness) on the best hypotheses and 96.4% (96.6%) on word graphs with 14.5 hypotheses per reference word. The system had a vocabulary of 5,336 words, used a bigram language model and was evaluated using a set of spontaneously spoken dialogues (EVAL96), that contained unknown words, hesitations and false starts.

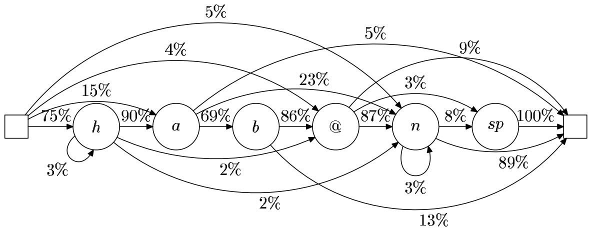
For our experiments, this system was used without modification as phone recognizer, using a simple monophone bigram “language model”. On a testset of 246 manually labeled utterances (MAN264), the recognizer achieved 68.0% phone accuracy (77.8% correct). All our experiments reported in the following sections are based on this recognizer.

We are currently tuning our recognizer for phone recognition (instead of word recognition). A significant increase in recognition accuracy has been achieved by using a different training method and a triphone bigram as language model; the accuracy increased to 75.5% (83.4%).

4.2. Lexical layer

The lexical layer of our prototype system employs discrete HMMs that emit phone symbols. For each syllable in our training corpus an HMM is constructed (one state per phone in the canonical pronunciation). Loops as well as forward skips are permitted and the emission probabilities are initialized using a phone-confusion matrix estimated from the recognition errors of our word-recognizer [2]. In the first experiment all syllable HMM states corresponding to the same phone (according to the canonical pronunciation) share one output distribution.

The model parameters are reestimated using a Baum-Welch training scheme. The training is performed on 12,000 utterances for which phone sequences were recognized by the first stage. Figure 3 shows the trained HMM for the word “haben” as an example.



state	symbol	prob.	state	symbol	prob.
h	h	0.91	@	@	0.90
	a	0.03		a	0.04
a:	a:	0.68	n	n	0.76
	a	0.28		m	0.22
b	b	0.85			
	m	0.11			

Figure 3: HMM for “haben”: transitions/emissions

With this approach the system achieved a word accuracy of 41.4% (48.2%) on our testset (not using a language model). This result is significantly worse than the result we achieved with our “conventional” decoder: 51.5% (58.7%). It should be noted that the two-step system is about twice as fast as the other one. In the 1996 Verbmobil evaluation our two-stage system achieved 66.6% accuracy on the best chain and 90.7% accuracy in a word graph (14.4 hypotheses per reference word)[1].

We experimented with various methods for training more specific lexical models. First, we trained separate output distributions for very frequent syllables (occurring more than 200, 300 and 500 times in our training corpus). We also used word-size HMMs to model very frequent words (see table 1).

We performed experiments in order to find the upper bound of the recognition rate that could possibly be

[1] These results were not obtained in the same evaluation category as those mentioned above.

threshold	syllables	words
200	41.8 (48.2)	42.5 (49.2)
300	42.0 (48.3)	42.7 (49.5)
500	41.8 (48.2)	42.7 (49.5)
baseline	41.4 (48.2)	
conventional	51.5 (58.7)	

Table 1: results for various degrees of sharing

achieved using the experimental model and to examine how much harder it is to classify the manually labeled phone transcriptions in contrast to classifying the canonically labeled data. We trained one set of models each for the canonical phone string (CAN246), the manually labeled data and the actual output of the phone recognizer (PHONE246) and then recognized the training set. For these tests the lexicon contained only the words that were in the training set. No language model was used. We achieved 96.6% word-accuracy on CAN246, 92.4% on MAN246 and 92.5% on PHONE246. It is surprising that the word error rate increases only about 4% from CAN246 to PHONE246 while the phone recognizer had introduced about 20-30% phone error rate to the input. These results can be explained by the fact that we used a phone-bigram "language model" which was trained on canonical pronunciations for the phone recognition thus we might have introduced regularities into the phone strings which were just predicted by the language model and not contained in the observation itself.

We would expect the manually labeled data to have less variation than the automatically derived transcriptions and thus the recognition performance on MAN246 should be better than on PHONE246. The observed recognition results may suggest that the current topology might not be adequate for all phenomena observed in the data. For instance, phone sequences can certainly not be adequately described by a first order Markov model. Therefore, we intend to use super-vectors to approximate a higher order Markov model.

5. RELATED WORK

Most attempts to produce better lexical models aim at finding the "optimal", the (weighted) n-best pronunciations or a pronunciation network for the set of words contained in a (large) phone labeled training corpus [1, 7, 9, 4, 6]. Alternatively phonetic rewriting rules can be learned from training data [5, 8, 3]. These can even be used to produce pronunciation hypotheses for words which are not in the training set. Both methods are used to replace a canonical lexicon by a new lexicon with trained pronunciations. Most approaches however do not feature explicit models that assign probabilities $\mathbf{P}(\mathbf{pron}|\mathbf{word})$ to any possible pronunciation of a given word.

6. CONCLUSIONS

We have presented a new approach to modelling pronunciation variations that frequently occur in spontaneous speech and pose a major problem for current speech recognition systems. Further work will concentrate on improved ("sharper") acoustical models, more adequate lexical HMM-architectures and the efficient coupling of the two layers to form an integrated model.

7. REFERENCES

[1] Andreas Hauenstein. *Aussprachewörterbücher zur automatischen Spracherkennung.* PhD thesis, Hamburg University, 1995.

[2] Kai Huebener, Uwe Jost, and Henrik Heine. Speech Recognition for Spontaneously Spoken German Dialogues. In *Proc. ICSLP 96*, Philadelphia, USA, 1996.

[3] J.J. Humphries, P.C. Woodland, and D. Pearce. Using Accent-specific Pronunciation Modelling for Robust Speech Recognition. In *Proc. ICSLP 96*, Philadelphia, USA, 1996.

[4] Frederick Jelinek. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–557, April 1976.

[5] Andreas Kipp, Maria-Barbara Wesenick, and Florian Schiel. Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora. In *Proc. ICSLP 96*, Philadelphia, USA, 1996.

[6] Michael D. Riley. A statistical model for generating pronunciation networks. In *Proceedings of the ICASSP91*, 1991.

[7] Tilo Sloboda and Alex Waibel. Dictionary Learning for Spontaneous Speech Recognition. In *Proc. ICSLP 96*, Philadelphia, USA, 1996.

[8] Maria-Barbara Wesenick. Automatic Generation of German Pronunciation Variants. In *Proc. ICSLP 96*, Philadelphia, USA, 1996.

[9] Christian-Michael Westendorf and Jens Jelitto. Learning Pronunciation Dictionary from Speech Data. In *Proc. ICSLP 96*, Philadelphia, USA, 1996.

[10] S.J. Young, N.H. Russell, and J.H.S. Thornton. Token passing: a simple conceptual model for connected speech recognition systems. Technical Report TR38, Cambridge University Engineering Department, July 1989.