PRONUNCIATION MODELING APPLIED TO AUTOMATIC SEGMENTATION OF SPONTANEOUS SPEECH

Andreas Kipp, Maria-Barbara Wesenick, Florian Schiel
IPSK University of Munich
kip|schiel|wesenick@phonetik.uni-muenchen.de

Abstract

In this paper[1] two different models of pronunciation are presented: the first model is based on a rule set compiled by an expert, while the second is statistically based, exploiting a survey about pronunciation variants occurring in training data. Both models generate pronunciation variants from the canonic forms of words. The two models are evaluated by applying them to the task of automatic segmentation of speech and then comparing the results to manual segmentations of the same speech data. Results show that correspondence between manual and automatic segmentations can be significantly improved if pronunciation variants are taken into account. The statistical model outperforms the rule based model.

1 Introduction

The modeling of pronunciation is becoming increasingly important in state-of-the-art ASR-systems. While allophonic variations of speech sounds can be modeled statistically by e.g. Hidden Markov Models or Artificial Neural Networks, a large number of possible pronunciation variants occurring in spontaneous speech extends beyond single speech sounds and reaches up to whole words or word tuples. Not even context-dependent acoustic models for sub-word units (like phonemes) are able to cover pronunciation variants of this kind. Therefore, pronunciation modeling is of great importance for many applications in speech technology.

In recent work [4] it has been shown that the consistent application of pronunciation variants for whole words and word tuples can improve the performance of an ASR-system. The approach, like others, involves the generation of pronunciation variants on the word level for the pronunciation lexicon component of an ASR-system.

The pronunciation models presented here are intended to be generic and not dependent on a specific lexicon. This is achieved by providing pronunciation variants for arbitrary phoneme sequences (*micro pronunciation variants*) and, in the case of the statistical pronunciation model, probabilities for the occurrence of a variant.

The phoneme sequences for which possible variants are provided can span word boundaries, because cross-word pronunciation variants occur frequently in spontaneous speech. The consideration of such variations led to improved results in various applications (e.g. [4]).

To show the ability to model the pronunciation of real speech data both models are applied in an automatic segmentation and labeling system for German spontaneous speech and compared with manual segmentations of the same data.

2 Pronunciation Modeling

Both pronunciation models discussed in this paper generate possible pronunciation variants given the reference transcription of an utterance. The reference transcription is the concatenation of the canonic forms of the words in the utterance. The canonic form is an arbitrary but unique phonemic transcription of a word spoken in isolation. The reference transcription of the orthographic representation of an utterance can therefore be determined by a simple lexicon lookup procedure. If an adequate inventory S of phoneme symbols (e.g. SAM-PA[2]) is used the reference transcription c can be denoted as a string of phoneme symbols $c = \gamma_0\gamma_1 \ldots \gamma_{N-1}$, $\gamma_i \in S$.

It is well known that in fluent speech the actual phonetic realizations of words often differ from the canonic form. This is especially true for spontaneous speech. The actual phonetic realization of an utterance will in most cases be different from its reference transcription. The actual realization r can be written as a (broad) phonetic transcription using the same inventory of phoneme symbols as the reference transcription, i.e. $r = \rho_0\rho_1 \ldots \rho_{M-1}$, $\rho_i \in S$.

To model the pronunciation of a reference transcription c the probability $p(r|c)$ for the occurrence of a certain realization r is stated. The structure of the model has to be suitable for ASR applications. Therefore, a first order Markov-chain which can be represented as a directed acyclic graph (DAG) was chosen as a model generating possible realizations for

[1] This work was funded by the German Federal Ministry for Research and Technology (BMBF) in the framework of the VERBMOBIL project.

[2] see `http://www.phon.ucl.ac.uk/home/sampa/home.htm`

a given reference transcription. The nodes of this DAG emit symbols from S and its edges specify possible transitions and their probabilities.

The DAG for a given c is constructed from micro pronunciation variants which specify possible alternative realizations for substrings of the reference transcription comprising a small number of phonemes (up to 10) and possibly spanning word boundaries. Formally, a micro pronunciation variant $\mathbf{m} \in M$ consists of a string a of symbols from S which can be substituted by a string b if a occurs in a certain context in the reference transcription. This pre- and post-context is specified by two strings x and y respectively. A micro pronunciation variant can be written as a tuple $\mathbf{m} = (x, a, y, b)$ of symbol string over S.

A micro pronunciation variant can be applied to c if it can be written as $c = sxayt$ where s is an arbitrary prefix and t is an arbitrary suffix of c. Note that the concatenation of strings a and b is denoted as ab and the length of a string a as $|a|$. The decomposition of c is not necessarily unique and therefore the location where a matches has to be considered. The set of all matching micro pronunciation variants and the corresponding location of the match is then given by

$$
\begin{aligned}
Q^{(c)} = \quad \{ \quad & (i, x, a, y, b) \bigm| (x, a, y, b) \in M \\
\wedge \quad & \gamma_{i-|x|} \dots \gamma_{i-1} = x \\
\wedge \quad & \gamma_i \dots \gamma_{i-|a|-1} = a \\
\wedge \quad & \gamma_{i+|a|} \dots \gamma_{i+|a|+|y|-1} = y \ \} \qquad (1)
\end{aligned}
$$

The DAG representing the Markov-chain contains the following elements:

- Nodes o_i emitting the reference transcription c. Each node o_i has the symbol $\gamma_i, i = 0 \dots N-1$ associated with it and has a transition to the node o_{i+1} (except for the last node $i = N-1$).
- For each $\mathbf{q_k} = (i, x, a, b), \mathbf{q_k} \in Q, k = 0 \dots |Q|$ a node or a node sequence $q_{k,j}$ emitting b. Each node $q_{k,j}$ has the symbol $\beta_j, j = 0 \dots |b|-1$ associated with it and, if $|b| > 1$ transitions to the successor node $q_{k,j+1}$ (for $j = 0 \dots |b|-2$).
- For each $\mathbf{q_k} = (i, x, a, b), \mathbf{q_k} \in Q$ transitions from the node o_{i-1} to $q_{k,0}$ and from the node $q_{k,|b|-1}$ to $o_{i+|a|}$.

The nodes o_i originate from the reference transcription and the nodes $q_{k,i}$ from pronunciation variants. Every path through the DAG from a initial node to an terminal node emits a possible pronunciation variant r for the given c. In a Markov-chain the overall probability of a symbol sequence, i.e $p(r|c)$ is the product over all transition probabilities along the path emitting the symbol sequence.

Both pronunciation models described below establish a DAG for a given reference transcription c. They differ in the set M of micro pronunciation variants and in the way transition probabilities are calculated.

2.1 Statistical Pronunciation Model

The set M of micro pronunciation variants is obtained by statistically evaluating a survey of pronunciation variants occurring in manually labeled training data.

For each utterance in the training corpus the reference transcription and the manually transcribed actual realization are subject to a maximum common subsequence alignment, yielding expressions for c and r of the form

$$
\begin{aligned}
c &= s_0 a_0 s_1 a_1 \dots a_L s_L \qquad (2) \\
r &= s_0 b_0 s_1 b_1 \dots b_L s_L \qquad (3)
\end{aligned}
$$

where the s_i are the common subsequences and each a_i in c has to be replaced by b_i to obtain r.

If a pre- and post-context of one symbol is considered and the s_i are written as the concatenation of their symbols $s_i = \sigma_{i,0} \dots \sigma_{i,|s_i|-1}$ each tuple $(\sigma_{i,|s_i|-1}, a_i, \sigma_{i+1,0}, b_i)$ can be considered as a micro pronunciation variant, and absolute counts $N_b(x, a, y, b)$ can be computed over the training corpus. This yields conditional probabilities $p_b(b|x, a, y)$ that the string a occurring in the context of x and y in a reference transcription is substituted by b if a substitution takes place (note that always $a_i \neq b_i$).

Additionally the probability $p_v(v|x, a, y) = 1 - p(\neg v|x, a, y)$ that substitution takes place at all has to be calculated (v denotes the event "substitution of a by b in the context of x and y"). This is done by relating the number of overall occurrences of the string xay, i.e. $N_v(x, a, y)$ with the number of occurrences where a replacement actually took place. This count is denoted by $N_v(v, x, a, y)$. Taking into account that

$$
\begin{aligned}
N_v(x, a, y) &= N_v(v, x, a, y) + \\
& \quad N_v(\neg v, x, a, y) \qquad (4) \\
N_v(v, x, a, y) &= \sum_{b \in \Theta} N_b(x, a, y, b) \qquad (5)
\end{aligned}
$$

where Θ is the set of all possible strings over S, simple maximum likelihood estimates can be given:

$$
\begin{aligned}
p_b(b|x, a, y) &= \frac{N_b(x, a, y, b)}{\sum_{\hat{b} \in \Theta} N_b(x, a, y, \hat{b})} \qquad (6) \\
p_v(v|x, a, y) &= \frac{N_v(v, x, a, y)}{N(x, a, y)} \qquad (7)
\end{aligned}
$$

Because training data are usually sparse discounting techniques well known from language modeling (e.g. [3]) can be applied to get more robust estimates.

The micro pronunciation variants observed in the training data establish the set M. Best results were obtained with a set of the size $|M|$ of approx. 1200, extracted from 72 dialogs (1245 turns) of *The Kiel Corpus of Spontaneous Speech* [1] dialog database.

A set $Q^{(c)}$ and a corresponding DAG for an arbitrary utterance with a reference transcription c can

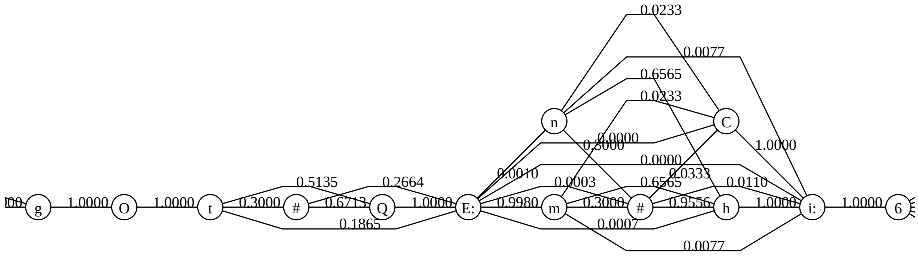

Figure 1: Part of a DAG representing pronunciation variants of an utterance. All edges are directed from left to right, the float values give transition probabilities.

now be given. For the calculation of the transition probabilities eqn. 6 and 7 have to be considered and moreover the fact that micro pronunciation variants in $Q^{(c)}$ might overlap and the application of one variant prevents other variants from being applied. In other words, statistical dependencies exist between micro pronunciation variants which have not been taken into account during training. However, a finite number of maximal non-overlapping subsets of $Q^{(c)}$ for which the assumption made during training is true can be found. A probability $p_c(\mathbf{q})$ which is the relative frequency of occurrences of $\mathbf{q}$ in these sets can be calculated and accounts for the context dependencies in a given c.

The probability for the application of $\mathbf{q_k} = (i, x, a, y, b)$ is equal to the probabilities $p(q_{k,0}) = \ldots = p(q_{k,|b|-1})$ that the symbol emitted by $q_{k,j}$ is in a realization r.

$$\begin{aligned} p(q_{k,0}) = \ldots = p(q_{k,|b|-1}) = \\ = p_c(\mathbf{q_k}) p_b(b|x,a,y) p_v(v|x,a,y) \quad (8) \\ \text{with:} \quad \mathbf{q_k} = (x,a,b,y), \quad \mathbf{q_k} \in Q^{(c)} \end{aligned}$$

The probability that no micro pronunciation variant is applied at index i is $p(o_i)$.

$$p(o_i) \;=\; 1 - \sum_{\hat{\mathbf{q}}_\mathbf{k} \in U_i} p_c(\hat{\mathbf{q}}_\mathbf{k}) p_b(b|x,a,y) p_v(v|x,a,b) \tag{9}$$

$$\text{with:} \quad \hat{\mathbf{q}}_\mathbf{k} = (x,a,b,y), \quad U_i \subset Q^{(c)}$$

where U_i is the subset of $Q^{(c)}$ that contains micro pronunciation variants spanning over the node o_i. With eqn. 8 and 9 the probabilities for all transitions contained in the graph can be expressed as:

$$p(q_{k,0}|o_{i-1}) = \frac{p(q_{k,0})}{p(o_i)} \tag{10}$$

$$p(o_{i+1}|o_i) = 1 - \sum_{k \in \Gamma^+(i)} p(q_{k,0}|o_i) \tag{11}$$

$$p(q_{k,i}|q_{k,i-1}) = 1 \tag{12}$$

$$p(o_i|q_{k,j}) = 1 \tag{13}$$

In eq. 11 o_i is a predecessor of all $q_{k,0}$ with $k \in \Gamma^+(i)$. The eqn. 10 through 13 state all transition probabilities occurring in the DAG according to the structural description given in section 2. Figure 1 shows an example.

2.2 Rule Based Pronunciation Model

This pronunciation model is based on a set of pronunciation rules compiled by a phonetician. These rules make up the set M. The rules were generated by evaluating a survey of pronunciation variants occurring in a speech database (PHONDAT II) and extrapolating the results to unseen but – from the phonetician's point of view – possible variants. At the moment the set comprises approx. 1500 rewrite rules. For a detailed description of the rule set see [5].

As there is no statistical information about the probabilities of rules, each variant contained in the resulting DAG is assumed to be equally likely and the transition probabilities are set accordingly.

3 Alignment

For an assessment of their ability to model the pronunciation of unseen speech data, DAGs produced by both pronunciation models were aligned to the corresponding speech signals containing the utterance. This alignment results in finding the transcription symbol sequence with the highest overall likelihood and a corresponding segmentation of the speech signal. A HTK [6] aligner with the following preprocessing settings and HMM-structure was used for this purpose:

- preprocessing with 13 MFCCs + first and second time derivative + Energy
- context independent phoneme models (SAMPA) with 3 to 5 states, 5 mixtures, no state-tying
- bootstrapping and isolated reestimation on a medium-size hand-labeled speech corpus

Best results for the statistical pronunciation model were obtained if the scores given by the transition probabilities in the graph were multiplied by a constant factor and incremented by a constant factor thereby giving more weight to the pronunciation modeling.

4 Evaluation and Results

For the model evaluation, segmentations produced with the HTK aligner as described above were compared with manual segmentations of the same data. This test data was excluded from the training data for the acoustic and the pronunciation model. In a comparison of two segmentations the accuracy in terms of the transcription and the segment boundaries was done after a longest common subsequence alignment between the segmentations concerned.

A fundamental problem lies in the fact that a unique correct segmentation and labeling of an utterance does not exist. Even carefully produced manual segmentations carried out by different individuals will differ from each other. Therefore, in addition to the comparison of manual and automatic segmentations, the manually produced ones were compared to each other [2].

	felix	marion	htkrla1	htkmr
dani	82.6	78.8	80.2	76.7
felix	-	79.9	80.3	77.2
marion	-	-	74.9	72.5

Table 1: Comparison between 3 manual segmentations (dani, felix, marion), an automatic segmentation with the statistical pronunciation model (htkrla1), and an automatic segmentation with the rule-based pronunciation model (htkmr).

Table 1 shows the symmetric accuracy[3] in terms of the transcription symbol sequence for one dialog of the VERBMOBIL corpus (approx. 5000 segments). Each cell gives the accuracy if the segmentation associated with the row is compared to that associated with the column.

The highest agreement exists among the two manual segmentations *dani* and *felix* but both differ considerably from the third manual segmentation *marion* and are even closer to the automatic segmentation produced with the statistical pronunciation model (*htkrla1*). The statistical pronunciation model consistently outperforms the rule-based model (*htkmr*) on this task.

In terms of accuracy of segment boundaries the comparison between manual segmentations shows a high agreement: on average 93% of all corresponding segment boundaries deviate less than 20ms from each other. The average percentage of corresponding segment boundary deviating less than 20ms in an automatic vs. a manual segmentation is 84%.

[3]The widely used accuracy measure $\frac{N-D-S-I}{N}$ relating the number of segments in the reference (N), deletions (D), substitutions (S) and insertions (I) which assumes that one of the segmentations is the reference is made symmetric by averaging with each segmentation once taken as a reference.

5 Conclusion

The results show that a high-quality segmentation and labeling can be generated if phonetic-phonological knowledge is used for modeling the pronunciation of spontaneous speech. This implies the usefulness of pronunciation modeling, especially statistically based, for other applications in speech technology.

The phonetic-phonological knowledge can be incorporated in the segmentation process by using a set of pronunciation rules or a statistical pronunciation model which is trained on data hand-labeled by phonetic experts. The former yields slightly worse performance but is independent of a specific domain. The latter leads to higher accuracy if test and training data are taken from the same domain.

The entropy of the graphs generated with the statistical pronunciation model is much lower than with the rule based model. This shows the close fitting to the domain and facilitates the task of the HMM aligner as the high accuracy of the resulting segmentation indicates. The lack of information in the case of the rule-based model, on the other hand, leads to very high entropy, i.e "ignorance" in the resulting graphs. ASR-applications using this kind of model therefore tend to make more errors.

Because of the promising results and its computationally efficient structure the statistical pronunciation model is at the moment being integrated into our HTK based speech recognizer.

References

[1] IPDS. The Kiel corpus of spontaneous speech. volume 1 + 2, Kiel, 1995.

[2] A. Kipp, M.-B. Wesenick, and F. Schiel. Automatic detection and segmentation of pronunciation variants in german speech corpora. In *Proceedings of the ICSLP*, volume 1, Philadelphia, oct. 1996.

[3] H. Ney and U. Essen. Estimating 'small' probabilities by leaving one out. In *Proceedings of the Eurospeech*, volume 3, pages 2239 – 2242, 1993.

[4] T. Sloboda and A. Waibel. Dictionary learnig for spontaneous speech. In *Proceedings of the ICSLP*, Philadelphia, oct. 1996.

[5] M.-B. Wesenick. Automatic generation of german pronunciation variants. In *Proceedings of the ICSLP*, volume 1, Philadelphia, oct. 1996.

[6] S. Young. *The HTK Book*. Cambridge University, 1995/1996.