TOWARDS DISCRIMINATIVE LEXICON OPTIMIZATION

Hauke Schramm *Peter Beyerlein*

Philips Research Laboratories
Weißhausstr. 2, D-52066 Aachen, Germany
e-mail: {Hauke.Schramm, Peter.Beyerlein}@philips.com

ABSTRACT

A lot of work has been done in deriving the pronunciation dictionary automatically from training data. These attempts focussed mainly on maximum likelihood or similar techniques. Due to the complexity and variability of the pronunciation process it is difficult to find an adequate pronunciation model. The model will deviate from the truth. Hence, the application of maximum likelihood techniques is likely to be suboptimal.
For this reason we present an approach, where the pronunciation model is learned discriminatively from data. The corresponding theory utilizes (1) probabilistic weighting of pronunciation variants of words and (2) discriminative model combination (DMC) based on Viterbi-approximations. We will show that the derived theory adjusts the weighting of pronunciation variants with respect to the word error rate, to the frequency of occurence of the specific pronunciation in the training data, and to the likelihood of the acoustic observation sequence given the pronunciation.

1. INTRODUCTION

In recent years recognition of conversational or spontaneous speech attracted increasing interest in the comunity. Due to the highly variable and less articulated utterances occuring in casual speech the ability of the pronunciation dictionary to cope with a wide range of pronunciation variation is of particular importance (see [1, 2, 3, 4, 5, 6], among others). Consequently many publications are concerned with the generation of pronunciation variants (e.g. [7, 8, 9]). It is however widely acknowledged that simply adding multiple pronunciations to the dictionary increases confusability between words. In many systems the integration of pronunciation variants is therefore accompanied with an appropriate incorporation of pronunciation prior probabilities into the decoding procedure [2, 3, 10]. Popular, due to its straightforward implementation, is the application of scaled, normalized unigram priors in combination with a maximum approximation over the pronunciation sequences. However although a sophisticated estimation procedure for determination of pronunciation priors is of substantial importance for an extended pronunciation modeling only few viable approaches exist. A straightforward method, apart from assuming equal distributions, is to estimate the priors by counting pronunciation frequencies observed on aligned training data. Other approaches employ decision trees or neural networks to predict surface pronunciations together with their likelihoods from canonical forms [9, 11]. This contribution aims at supplementing former approaches by an extended pronunciation prior estimation technique that employs discriminative training methods. A related method is presented in [12], where the pronunciation network is constructed and trained with respect to the sentence error using the MCE approach. Our approach is based on an extension of the discriminative model combination (DMC) ([13], [14]) by fitting the unigram priors of pronunciation alternatives into the DMC-framework. Here the training objective is directly related to the word error rate on training data. In the following sections, we discuss the integration of multiple pronunciations in the decoding procedure and give a short overview of DMC. In section 4, we develop the new theory.

2. INTEGRATION OF MULTIPLE PRONUNCIATIONS IN A DECODER

In the sequel w will be associated with a unique linguistic word entity. The wordlist $\Delta = w_1, w_2, ..., w_D$ of our dictionary has D entries, each , w_l, having a certain number V_l of individual pronunciations $v_{l1}, v_{l2}, ..., v_{lV_l}$. The unigram probability distribution of w_l which is given by $p(v_{lj}|w_l)$ is subjected to the normalization constraint

$$\sum_{j=1}^{V_l} p(v_{lj}|w_l) = 1, \quad V_l \geq 1, \quad \forall w_l$$

In our approach multiple pronunciations of the same linguistic word are tied with respect to the m-gram language models, i.e.

$$p(v_{lj} \mid H) = p(w_l|H), \quad j = 1, ..., V_l, \quad \forall w_l,$$

with $H = \{w_1, w_2, ..., w_{m-1}\}$ being the m-gram word history preceding w_l. Now let $w_1^N = w_1, w_2, ..., w_N$ be a hypothesized word sequence and $v_1^N = v_1, v_2, ..., v_N$ be one possible sequence of pronunciations related to w_1^N. By $x_1^T = x_1, x_2, ..., x_T$ we will denote the input feature vectors and by $s_1^T = s_1, s_2, ..., s_T$ a state sequence trajectory through the concatenated pronunciation HMMs. According to the decision rule the decoded word sequence $\hat{w}_1^N$ is defined as

$$\hat{w}_1^N = \arg\max_{w_1^N} p(w_1^N|x_1^T) \qquad (1)$$

Explicite integration of alternate pronunciation sequences is obtained by

$$\begin{aligned}\hat{w}_1^N &= \arg\max_{w_1^N} \sum_{v_1^N \epsilon \Upsilon(w_1^N)} p(w_1^N, v_1^N | x_1^T) \\ &= \arg\max_{w_1^N} \sum_{v_1^N \epsilon \Upsilon(w_1^N)} \sum_{s_1^T \epsilon \Psi(v_1^N)} p(w_1^N, v_1^N, s_1^T | x_1^T)\end{aligned} \tag{2}$$

where $\Upsilon(w_1^N)$ denotes the set of v_1^N linguistically equivalent to w_1^N and $\Psi(v_1^N)$ denotes the set of trajectories s_1^T through the compound HMM of v_1^N. Factorizing (2) into products of individual word contributions and restricting the pronunciation priors to unigram probabilities:

$$p(v_1^N | w_1^N) = \prod_{i=1}^{N} p(v_i | w_i) \tag{3}$$

leads to the following optimization problem:

$$\hat{w}_1^N = \arg\max_{w_1^N} \left\{ p(w_1^N) \cdot \sum_{v_1^N \epsilon \Upsilon(w_1^N)} \left[\prod_{i=1}^{N} p(v_i | w_i) \right] \cdot p(x_1^T | v_1^N) \right\}. \tag{4}$$

Usually the sum over all v_1^N is approximated by the most likely pronunciation sequence. However approximating the sum time-synchronously has been shown to give small but significant improvement [15].

3. DISCRIMINATIVE MODEL COMBINATION

3.1. Log-linear Model Combinations

Applying a multi-pass decoding strategy is typically the way to incorporate multiple model sets into the decoder of state-of-the-art speech recognition systems which use multiple acoustic and language model sets. A simpler and still optimal alternative to sophisticated multiple pass decoding strategies was used successfully by Philips in the 1998 Hub4 evaluation: Discriminative model combination (DMC) [6]. DMC integrates all available models into one decoding pass. Thus we acquire a decoder containing information combined directly from all model sets.

For the sake of simplicity we now introduce k as abbreviation for w_1^N and x for x_1^T. Assuming we are given M different models of any kind, numbered $j = 1, \ldots, M$. From model j we can compute the posterior probability $p_j(k|x), p_j(k'|x)$ of hypothesized classes k, k' given an observation x. The models are now log-linearly combined into a distribution of the exponential family:

$$p_{\Lambda}(k|x) = e^{-\log Z_{\Lambda}(x) + \sum_{j=1}^{M} \lambda_j \log p_j(k|x)} \tag{5}$$

The coefficients $\Lambda = (\lambda_1, \ldots, \lambda_M)^T$ can be interpreted as weights of the models j within the model combination (5). The value $Z_{\Lambda}(x)$ is a normalization constant. As opposed to the maximum entropy approach, which leads to a distribution of the same functional form, the coefficients Λ are optimized with respect to the word error rate of the discriminant function (6):

$$\log \frac{p_{\Lambda}(k|x)}{p_{\Lambda}(k'|x)} = \sum_{j=1}^{M} \lambda_j \log \frac{p_j(k|x)}{p_j(k'|x)} \tag{6}$$

DMC will optimize the so called language weight (or language model factor), if only one acoustic and one language model are combined. Now, since the weight λ_j of the model j within the combination depends on its ability to provide information for correct classification, DMC allows for the optimal integration of any set of models into one decoder. In this contribution we will represent the pronunciation weights (3) in a log-linear functional form such that the DMC framework may be applied.

3.2. Minimum Word Error Training

We are given a set of sentences $n = 1, \ldots, H$ for DMC training . For every training sentence we observe x_n (spoken utterance) and we know the correct class assignment k_n (spoken word sequence). We can define the set of rival classes $k \neq k_n$ using a preliminary decoding (if appropriate), and the number of word errors of the rival class k can be computed with the help of the Levenstein distance $\Gamma(k_n, k)$. The model combination should then minimize the word error count $E(\Lambda)$

$$E(\Lambda) = \sum_{n=1}^{H} \Gamma\left(k_n, \arg\max_{k} \left(\log \frac{p_{\Lambda}(k|x_n)}{p_{\Lambda}(k_n|x_n)}\right)\right) \tag{7}$$

on representative training data to assure optimality on an independent test set. As this optimization criterion is not differentiable, we approximate it by a smoothed word error count:

$$E_S(\Lambda) = \sum_{n=1}^{H} \sum_{k \neq k_n} \Gamma(k, k_n) S(k, n, \Lambda), \tag{8}$$

where $S(k, n, \Lambda)$ is a smoothed indicator function. If the classifier (6) selects hypothesis k, $S(k, n, \Lambda)$ should be close to one, and if the classifier rejects hypothesis k, it should be close to zero. A possible indicator function with these properties is

$$S(k, n, \Lambda) = \frac{p_{\Lambda}(k|x_n)^{\eta}}{\sum_{k'} p_{\Lambda}(k'|x_n)^{\eta}}, \tag{9}$$

where η is a suitable constant. An iterative gradient descent scheme is obtained from the optimization of $E_S(\Lambda)$ with respect to Λ:

$$\lambda_j^{(0)} = 0 \quad \text{(Uniform Distribution)}$$

$$\begin{aligned}
\lambda_j^{(I+1)} &= \lambda_j^{(I)} - \frac{\varepsilon \cdot \eta}{\sum_{n=1}^{H} L_n} \sum_{n=1}^{H} \sum_{k \neq k_n} S(k,n,\Lambda^{(I)}) \cdot \\
&\quad \cdot \tilde{\Gamma}(k,n,\Lambda^{(I)}) \cdot \log \frac{p_j(k|x_n)}{p_j(k_n|x_n)} \\
\Lambda^{(I)} &= (\lambda_1^{(I)}, \ldots, \lambda_M^{(I)})^T \qquad (10) \\
&\quad j = 1, \ldots, M \\
\tilde{\Gamma}(k,n,\Lambda) &= \Gamma(k,k_n) - \sum_{k' \neq k_n} S(k',n,\Lambda)\Gamma(k',k_n).
\end{aligned}$$

with L_n being the number of words in sentence n.

4. DISCRIMINATIVE WEIGHTING OF PRONUNCIATION VARIANTS

The outline of this section is as follows: Starting out from a log-linear combination of acoustic and language model as described in the preceding section we will introduce multiple pronunciation sequences into the model combination (5). Defining an appropriate modeling of pronunciation weights leads us to a DMC-model definition p_j that associates each DMC-weight λ_j to the weight of a distinct pronunciation. By discriminatively optimizing the DMC-weights we therefore optimize the pronunciation variant weights.

We now introduce multiple pronunciation sequences into the model combination equation by using the following correspondence derived in section 2

$$p(x_1^T|w_1^N) = \sum_{v_1^N \epsilon \Upsilon(w_1^N)} \left[\prod_{i=1}^{N} p(v_i|w_i) \right] \cdot p(x_1^T|v_1^N). \qquad (11)$$

Applying the maximum approximation and denoting by $\tilde{v}_1^N = \tilde{v}_1, \tilde{v}_2, ..., \tilde{v}_N$ the most likely pronunciation sequence

$$\tilde{v}_1^N = \arg \max_{v_1^N \epsilon \Upsilon(w_1^N)} \left[\prod_{i=1}^{N} p(v_i|w_i) \right] \cdot p(x_1^T|v_1^N) \qquad (12)$$

we obtain

$$\log p(x_1^T|w_1^N) \overset{max}{=} \log \prod_{i=1}^{N} p(\tilde{v}_i|w_i) + \log p(x_1^T|\tilde{v}_1^N) \qquad (13)$$

The next step is to substitute the sentence based notation of the first term by a dictionary based one. Let $h_{lj}(\tilde{v}_1^N)$ be the frequency of occurrence of variant v_{lj} in $\tilde{v}_1^N$ (the value v_{lj} was defined in section 2). Then we obtain

$$\begin{aligned}
&\log p(x_1^T|w_1^N) \overset{max}{=} \\
&\qquad \log \prod_{l=1}^{D} \prod_{j=1}^{V_l} p(v_{lj}|w_l)^{h_{lj}(\tilde{v}_1^N)} + \log p(x_1^T|\tilde{v}_1^N) \qquad (14)
\end{aligned}$$

As $p(v_{lj}|w_l)$ has a scalar value we can rewrite it as

$$p(v_{lj}|w_l) = e^{\lambda_{lj}}, \quad \lambda_{lj} \leq 0 \qquad (15)$$

Adjusting $p(v_{lj}|w_l)$ is now done only by modifying λ_{lj} which initial value could be given by the maximum likelihood estimate obtained by frequency counting of the respective pronunciation variant. Inserting this into (14), after reformulation, we obtain

$$\log p(x_1^T|w_1^N) \overset{max}{=} \sum_{l=1}^{D} \sum_{j=1}^{V_l} \lambda_{lj} h_{lj}(\tilde{v}_1^N) + \log p(x_1^T|\tilde{v}_1^N) \quad (16)$$

For the model combination equation (5) we obtain

$$p_\Lambda(w_1^N|x_1^T) = \frac{h(w_1^N, \tilde{v}_1^N, x_1^T, \Lambda)}{\sum_{w'{}_1^{N'}} h(w'{}_1^{N'}, \tilde{v}'{}_1^{N'}, x_1^T, \Lambda)} \qquad (17)$$

with

$$\begin{aligned}
\log h(w_1^N, \tilde{v}_1^N, x_1^T, \Lambda) = \;& \lambda_1 \log p(w_1^N) + \lambda_2 \log p(x_1^T|\tilde{v}_1^N) \\
&+ \sum_{l=1}^{D} \sum_{j=1}^{V_l} \lambda_{lj} h_{lj}(\tilde{v}_1^N) \qquad (18)
\end{aligned}$$

Thus we arrive at a log-linear combination of the language model $p(w_1^N)$, the acoustic model $p(x_1^T|\tilde{v}_1^N)$ and the $\sum_{l=1}^{D} \sum_{j=1}^{V_l}$ pronunciation frequency models $h_{lj}(\tilde{v}_1^N)$.

Minimization Of The Smoothed Word Error Rate Application of the gradient descent scheme (10) to the model combination (17), (18) leads to following iteration scheme:

$$\begin{aligned}
\lambda_{lj}^{(I+1)} &= \lambda_{lj}^{(I)} - \frac{\epsilon \cdot \eta}{\sum_{n=1}^{H} L_n} \sum_{n=1}^{H} \sum_{k \neq k_n} S(k,n,\Lambda^{(I)}) \cdot \\
&\quad \tilde{\Gamma}(k,n,\Lambda^{(I)}) \cdot (h_{lj}(\tilde{v}(k)) - h_{lj}(\tilde{v}(k_n))), \\
&\qquad (19)
\end{aligned}$$

where $\tilde{v}(k)$ denotes the optimal pronunciation variant sequence, which corresponds to the word sequence k given the acoustic utterance x_1^T.

Note that the adjustment of pronunciation weights is especially influenced by

- the error term $\tilde{\Gamma}(k,n,\Lambda)$
- the frequency of occurence of the pronunciation v_{lj} in the true word sequence k_n and the rival k: $(h_{lj}(\tilde{v}(k)) - h_{lj}(\tilde{v}(k_n)))$ and
- the underlying acoustic models via the smoothing function $S()$

From equation (19) we see that for good hypothesis ($\tilde{\Gamma}(k, n, \Lambda) < 0$) we should

- increase $p(v_{lj}|w_l)$, if $h_{lj}(\tilde{v}(k)) > h_{lj}(\tilde{v}(k_n))$ and
- decrease $p(v_{lj}|w_l)$, if $h_{lj}(\tilde{v}(k)) < h_{lj}(\tilde{v}(k_n))$.

This means we should increase the weight of variants which are particularly frequent in successful hypotheses while reducing the weight of particularly rare variants. With a similar argumentation it can be shown that it is the other way around for unsuccessful hypotheses. Note the difference to the maximum likelihood solution where the relative frequency of occurrence of a specific variant in the *correct* hypotheses determines its pronunciation unigram prior.

5. BUILDING THE PRONUNCIATION DICTIONARY

Let us assume that we are given a (large) set of pronunciation variants for each word in the vocabulary. The described method can then be used in an iterative manner by (1) optimizing the pronunciation weights and (2) removing entries with a weight below a given threshold. By this, a discriminative adaptation of a large background dictionary to new tasks, dialects or speakers could be possible. By introducing different HMM topologies for each pronunciation the method could moreover be used to realize a restricted discriminative HMM topology optimization.

6. SUMMARY

We presented a new approach, where the pronunciation model is optimized with respect to the word error rate using the DMC framework. The obtained solution utilizes the word error rate, the frequency of occurence of a pronunciation and the likelihood of the acoustic observation sequence to compute the pronunciation priors of the lexical model.

7. REFERENCES

1. Michael D. Riley, “A Statistical Model For Generating Pronunciation Networks” in Proc. ICASSP, pages 737–740, Toronto, USA, May 1991.
2. B. Peskin, M. Newman, Don McAllaster, ”Improvements In Recognition Of Conversational Telephone Speech” In Proc. EUROSPEECH’97, Rhodes, Greece, Sep. pages 22–25, 1997.
3. S. Wegmann, P. Zhan, I. Carp, M. Newman, J. Yamron, and L. Gillick ”Dragon Systems’ 1998 Broadcast News Transcription System”, In Proc. DARPA Broadcast News and Transcription Workshop, Herndon, Virginia, Feb. 1999.
4. Martine Adda-Decker and Lori Lamel, ”Pronunciation Variants Across Systems, Languages and Speaking Style” In Proc. ESCA Workshop “Modeling Pronunciation Variation For Automatic Speech Recognition”, Rodulc, May 98.
5. K. Beulen, S. Ortmanns, A. Eiden, S. Martin, L. Welling, J. Overmann, H. Ney, ”Pronunciation Modeling In The RWTH Large Vocabulary Speech Recognizer”, In Proc. of the ESCA Workshop ’Modeling Pronunciation Variation For Automatic Speech Recognition’, Rodulc, May 1998
6. P. Beyerlein, X. L. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A. Wendmuth, S. Molau, M. Pitz, A. Sixtus, ”The Philips/RWTH System for Transcription of Broadcast News”, In Proc. DARPA Broadcast News and Transcription Workshop, Herndon, Virginia, Feb. 1999.
7. T. Sloboda, A. Waibel “Dictionary Learning For Spontaneous Speech Recognition” Proc. Int. Conf. on Spoken Language Processing, Philadelphia, PA, September 1996
8. Byrne, W., Finke, M., Khudanpur, S., McDonough, J., Nock, H., Riley, M., Saraclar, M., Wooters, C., Zavaliagkos, G. “Pronunciation modelling for conversational speech recognition: A status report from WS97”. In IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings (ASRU), Santa Barbara, CA, USA, pp. 26–33.
9. Fosler-Lussier, E. “Multi-level decision trees for static and dynamic pronunciation models” Proc. of the European Conference on Speech Communication and Technology (Eurospeech), Budapest, Hungary, pp. 463–466.
10. Hochberg, M. M., Renals S. J. and Robinson A. J. “ABBOT: The CUED hybrid connectionist-HMM large-vocabulary recognition system” in Proc. of Spoken Language Systems Technology Workshop, ARPA, March 1994.
11. T. Fukada, T. Yoshimura and Y. Sagisaka, ”Automatic Generation Of Multiple Pronunciations Based On Neural Networks And Language Statistic”, In Proc. of the ESCA Workshop ’Modeling Pronunciation Variation For Automatic Speech Recognition’, Rodulc, May 1998
12. F. Korkmazskiy, B.-H. Juang “Discriminative Training of the Pronunciation Networks”, Proc. Automatic Speech Recognition and Understanding, Santa Barbara, pp. 223–229, 1997
13. P. Beyerlein, “Diskriminative Modellkombination in Spracherkennungssystemen mit gro”sem Wortschatz”, Dissertation, Lehrstuhl für Informatik VI, RWTH Aachen, 1999
14. P. Beyerlein, “Discriminative Model Combination”, in Proceedings of 1997 IEEE Workshop on Automatic Speech Recognition and Understanding, Santa Barbara, pp. 238-245, Dec. 1997.
15. Schramm, H. and Aubert, X., ”Efficient Integration Of Multiple Pronunciations In A Large Vocabulary Decoder”, Proc. ICASSP, Istanbul, Turkey, June 2000.