RATIONAL INTERPOLATION OF MAXIMUM LIKELIHOOD PREDICTORS IN STOCHASTIC LANGUAGE MODELING

Ernst Günter Schukat-Talamazzini[1], *Florian Gallwitz*[2], *Stefan Harbeck*[2], *Volker Warnke*[2]

[1]Institute for Computer Science
University of Jena, Germany
Ernst-Abbe-Platz 1-4
D-07740 Jena, Germany
e-mail: schukat@informatik.uni-jena.de

[2]Chair for Pattern Recognition
University of Erlangen-Nuremberg
Martensstrasse 3
D-91058 Erlangen, Germany
e-mail: {*name*}@informatik.uni-erlangen.de

ABSTRACT

In our paper, we address the problem of estimating stochastic language models based on n-gram statistics. We present a novel approach, *rational interpolation*, for the combination of a competing set of conditional n-gram word probability predictors, which consistently outperforms the traditional linear interpolation scheme. The superiority of rational interpolation is substantiated by experimental results from language modeling, speech recognition, dialog act classification, and language identification.

1. INTRODUCTION

In our paper, we address the problem of estimating stochastic language models $P(w)$ for sentences $w = w_1 \ldots w_T$ of words w_t from a finite vocabulary $\mathcal{V}$. The joint distribution $P(w)$ can be decomposed by the well-known chain rule

$$P(w) \;=\; \prod_{t=1}^{T} P(w_t|w_1^{t-1}) \;=\; \prod_{t=1}^{T} P(w_t \mid w_1 \ldots w_{t-1}) \quad (1)$$

into a product of conditional word probabilities (by w_s^t we denote the substring $w_s \ldots w_t$ of w). The latter, in turn, are usually approximated by conditional bigram or conditional trigram probabilities [5] or are evaluated without explicit history pruning as in the polygram model [16, 8] or Bell Lab's variable n-gram stochastic automata [14].

It is straightforward to replace the conditional n-gram probabilities on the right hand side of Eq. (1) by their maximum likelihood estimates

$$\hat{P}(w_t|w_{t-n+1}^{t-1}) \;=\; \frac{\#(w_{t-n+1}^{t})}{\#(w_{t-n+1}^{t-1})} \quad (2)$$

where the function $\#(\cdot)$ counts the frequency of occurrence of its argument word sequence in the training corpus. Unfortunately, due to the sparse data problem the frequency ratios of Eq. (2) are far from being reliable probability estimates even in case of moderate order n. In particular, the quantity $\hat{P}(w_t|w_{t-n+1}^{t-1})$ degenerates to zero if the n-gram w_{t-n+1}^{t} was never observed in the training data, and what is more: it becomes undefined as soon as the denominator $\#(w_{t-n+1}^{t-1})$ of the MLE expression turns to zero.

As a consequence, the raw ML estimates have to be smoothed by a suitable backing-off or interpolation strategy. Backing-off approaches such as Katz' trigram formula [7] typically operate on a discounted version of n-gram frequencies, for example based on Jeffrey's rule or the Good-Turing estimate [2] of unseen events, reducing the occurrence counts of frequent events in favour of the rare ones. The probability mass that was saved by deriving the conditional trigram probabilities from discounted frequencies is then redistributed to the unseen trigram events according to a lower order (in this case: bigram) language model.

Katz' formula handles the statistics of unseen events and leaves the remaining estimates essentially unchanged. In order to fully exploit the information represented in the lower order models, the competing ML estimates should better be combined using a linear interpolation scheme like the smoothed trigram model [5]

$$\tilde{P}(w|uv) \;=\; \lambda_0\frac{1}{L}+\lambda_1\hat{P}(w)+\lambda_2\hat{P}(w|v)+\lambda_3\hat{P}(w|uv) \quad (3)$$

The weights λ_i of this convex combination of conditional trigram, bigram, unigram, and zerogram ($1/L$) probabilities can be optimized with respect to the maximum likelihood of an independent cross-validation data set by running an instance of the well-known EM algorithm [3].

Both backing-off and linear interpolation do not precisely distinguish between *reliable* predictors (i.e.: the relevant word history has a reasonably large # value) and less reliable predictors. Thus, in numerous situations the contributions of ML predictors $\hat{P}(w|v)$ will not be weighted in an optimal way. Several authors proposed alternative interpolation schemes [10, 13] incorporating certain predictor weights into the model Eq. (3) which are expected to appropriately reflect our confidence in the component probability estimators.

These models, however, abandoned the use of interpolation along with its benefit of (cross-validation-) data-driven adaptation of the smoothing process. In this paper we present a novel approach of predictor fusion, *rational interpolation*, combining the profits of linear interpolation with the introduction of an explicit reliability scoring into the model. Rational interpolation is shown to consistently outperform the linear one in the realm of language modeling and, beyond that, is a promising competitor to EM-based deleted interpolation [6] when tackling smoothing problems in decision tree design and acoustic modeling (interpolation of context-dependent phone HMMs).

The remainder of this paper is organized as follows: Section 2 describes our rational probability model as well as a gradient ascent algorithm designed to optimize its interpolation coefficients. In Section 3 a particular weighting function of hyperbolic shape is introduced that serves to score the language model predictors. Finally, Section 4 and Section 5 present test set perplexities of linear and rational models achieved on different text corpora as well as results from application of our novel interpolation technique in speech recognition, dialog act classification, and language identification; Section 6 concludes the paper.

2. INTERPOLATION OF WEIGHTED PREDICTORS

In general, we consider *polygram* models of the form

$$\tilde{P}(w|v) \;=\; \sum_{i\in I} \lambda_i \cdot \hat{P}_i(w|v), \quad (4)$$

where the $\hat{P}_i$ predict the actual word w based on some appropriate portion of the sentence history v. Consider, for example, the usual ML estimators of conditional n-gram

 © E.G. Schukat-Talamazzini 1997

probabilities (Eq. (2)) as a possible choice for the $\hat{P}_i$'s; similarly, predictors may be conceived as based on non-contiguous statistics of the word history like distance-τ bigrams ([11, 15], see also Section 4).

2.1. Rational Interpolation

Now let us introduce a history-dependent weight function $g_i(v)$, scoring the predictor reliability based on that portion of v which is relevant to $\hat{P}_i$, for instance the past bigram uv for the trigram predictor $\hat{P}(w|uv)$ (see Section 3 for more details on g_i). Linear interpolation of the weighted predictors leads to the expression

$$P'(w|v) \stackrel{\text{def}}{=} \sum_{i \in I} \lambda_i \cdot g_i(v) \cdot \hat{P}_i(w|v) \qquad (5)$$

Due to the violation of the normality condition, $P'(w|v)$ does not yet represent a probability distribution. After division by the appropriate renormalization factor

$$P''(v) \stackrel{\text{def}}{=} \sum_{w \in \mathcal{V}} P'(w|v) = \sum_{i \in I} \lambda_i \cdot g_i(v) \qquad (6)$$

we obtain the rational interpolation model

$$\tilde{P}(w|v) \stackrel{\text{def}}{=} \frac{P'(w|v)}{P''(v)} = \frac{\sum_{i \in I} \lambda_i \cdot g_i(v) \cdot \hat{P}_i(w|v)}{\sum_{i \in I} \lambda_i \cdot g_i(v)} \qquad (7)$$

which is no longer linear in its interpolation coefficients.

2.2. Optimizing the Interpolation Coefficients

The optimization of the λ_i makes use of a cross-validation data set $w = w_1 \ldots w_S$; let v_s denote the past sentence context of word item w_s, $s = 1, \ldots, S$. The desired coefficient vector λ results from maximizing the log likelihood function

$$\ell_w(\lambda) \stackrel{\text{def}}{=} \log \tilde{P}(w) = \log \prod_{s=1}^{S} \tilde{P}(w_s|w_1^{s-1}) \qquad (8)$$

of the validation data w.r.t. λ. Note that in contrast to the linear scheme Eq. (4) the rational model Eq. (7) cannot be interpreted as a doubly stochastic process; consequently, the EM algorithm [3, 6] is not applicable in order to estimate the λ_i's. However, since the usual normality condition $\sum_i \lambda_i = 1$ factors out in Eq. (7), the λ_i may be iteratively optimized by unconstrained gradient ascent using the gradient vector $\nabla_\ell(\lambda)$ with components

$$\frac{\partial \ell}{\partial \lambda_i} = \sum_{s=1}^{S} \left\{ \frac{g_i(v_s) \cdot \hat{P}_i(w_s|v_s)}{P'(w_s|v_s)} - \frac{g_i(v_s)}{P''(v_s)} \right\} \qquad (9)$$

The Hessian matrix $H = H(\lambda)$, $H_{ij} = \partial^2 \ell \,/\, \partial\lambda_i \partial\lambda_j$ has the form $H = H' - H''$ where the matrices H', H'' with elements

$$H'_{ij} = \sum_{s=1}^{S} \frac{g_i(v_s) \cdot \hat{P}_i(w_s|v_s) \cdot g_j(v_s) \cdot \hat{P}_j(w_s|v_s)}{(P'(w_s|v_s))^2}$$

$$H''_{ij} = \sum_{s=1}^{S} \frac{g_i(v_s) \cdot g_j(v_s)}{(P''(v_s))^2} \qquad (10)$$

are symmetric and positive-definite. Unfortunately, $H(\lambda)$ itself is in general *not* positive-definite, and the Newton iteration

$$\lambda^{(k+1)} = \lambda^{(k)} + (H(\lambda^{(k)}))^{-1} \cdot \nabla_\ell(\lambda^{(k)}) \qquad (11)$$

is not applicable here since the eigenvalues of H may turn to zero (no inverse exists!) or turn negative; in the latter case, $\lambda^{(k+1)}$ is no longer guaranteed to increase the likelihood function $\ell_w(\cdot)$. In our experiments we ran the gradient ascent shown in Eq. (11), replacing the original Hessian H by H' which resulted in a fast converging coefficient set with monotonical improvement of $\ell_w(\cdot)$ for all data sets under test.

3. HYPERBOLIC WEIGHT FUNCTIONS

What remains to be done is the specification of our context-dependent reliability scores $g_i(v)$. Recall that we assumed the predictors to be of the form of relative frequency estimators

$$\hat{P}_i(w|v) = \frac{\#_i(v,w)}{\sum_u \#_i(v,u)} = \frac{\#_i(v,w)}{\#_i(v)} \qquad (12)$$

with appropriately defined marginals $\#_i(\cdot)$ of the basic occurrence counts $\#(\cdot)$. It is reasonable to rate our confidence in the estimator $\hat{P}_i(w|v)$ by a function g_i which is monotonically increasing with the absolute number $\#_i(v)$ of events this relative frequency is based upon. Choosing a hyperbolic weight function

$$g_i(v) = \frac{\#_i(v)}{\#_i(v) + C} \qquad (13)$$

(with $C > 0$), we obtain the rational interpolation model

$$\tilde{P}(w|v) = \frac{\sum_{i \in I} \lambda_i \cdot \frac{\#_i(v,w)}{\#_i(v) + C}}{\sum_{i \in I} \lambda_i \cdot \frac{\#_i(v)}{\#_i(v) + C}} \qquad (14)$$

If the constant C tends to zero, we get $g_i(v) \equiv 1$ and Eq. (14) reduces to the well-known linear interpolation of conditional n-gram probabilities; whereas if C approaches infinity, $\tilde{P}_i(w|v)$ is computed as the ratio of linearly interpolated marginals of the statistics $\#(v,w)$ and $\#(v)$:

$$\tilde{P}(w|v) = \sum_{i \in I} \lambda_i \cdot \#_i(v,w) \;/\; \sum_{i \in I} \lambda_i \cdot \#_i(v) \qquad (15)$$

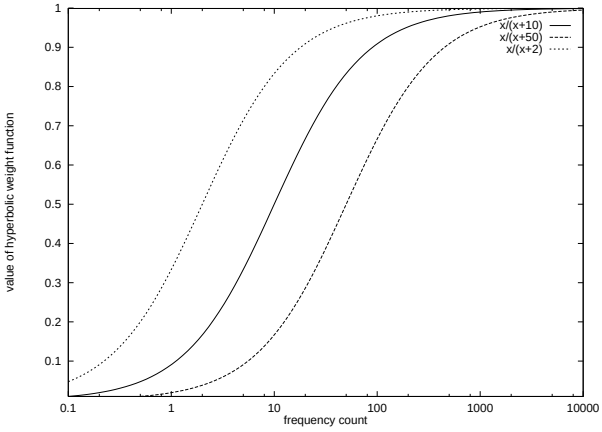

Figure 1. Hyperbolic weight functions ($C = 2, 10, 50$)

Generally, the values $g_i(v)$ vary inside the unit interval (see Fig. 1) and serve to emphasize estimates which produce their word predictions from contexts with safe statistics. Note that in the rational language model with hyperbolic score g_i we don't have to worry about the zero frequency problem, since all denominators in Eq. (14) are strictly positive if only $C > 0$ and at least one marginal context count $\#_i(v)$ is nonzero. A second advantageous

property of rational models that shall be mentioned here is the smooth dependence of predictor contribution on the frequency of the relevant word context, which was introduced by our hyperbolic weight function; as a consequence, there is no more need in a rational model artificially to introduce dependences of the interpolation coefficients λ_i from the word history v ("bucketing").

4. EXPERIMENTAL RESULTS I

In order to assess the performance of our rational interpolation approach we tested language models with three different sets of predictors; each of these models in turn is scaled by the maximum order n of context considered in estimating the conditional word probabilities ($n = 2, \ldots, 6$):

- [poly]: the n^{th} order polygram model interpolates the k-gram predictors $\hat{P}_k = \hat{P}(w_t|w_{t-k+1}^{t-1})$ for $k = 0, \ldots, n$. The poly predictor set of order 3, for instance, results in the classical trigram model of Eq. (3).
- [poly+2]: this model contains the above k-gram predictors augmented by those distance-τ conditional bigrams $\hat{P}_{2/\tau} = \hat{P}(w_t|w_{t-\tau})$ falling into the given maximum context, i.e. $\tau = 1, \ldots, n-1$.
- [poly+3]: in excess to k-gram and distance-τ bigram predictors, this model incorporates additional $^{(n-1)(n-2)}/_2$ distance-τ,σ conditional trigrams $\hat{P}_{3/\tau,\sigma} = \hat{P}(w_t|w_{t-\tau-\sigma}, w_{t-\tau})$, the gap parameter τ ranging from 1 to $n-2$ and σ ranging from 1 to $n-\tau-1$.

	$k = ?$				$\tau = ?$				$^{\tau}/_{\sigma} = ?$					
	2	3	4	5	1	2	3	4	$^1/_1$	$^1/_2$	$^1/_3$	$^2/_1$	$^2/_2$	$^3/_1$
w_{t-1}	●	●	●	●	●	○	○	○	●	●	●	○	○	○
w_{t-2}	○	●	●	●	○	●	○	○	●	○	○	●	●	○
w_{t-3}	○	○	●	●	○	○	●	○	○	●	○	●	○	●
w_{t-4}	○	○	○	●	○	○	○	●	○	○	●	○	●	●
	k-grams				τ-bigrams				τ,σ-trigrams					

Figure 2. Predictors of a poly+3 model of order 5

The filled circles in Fig. 2 indicate exactly those word history positions w_{t-4}, w_{t-3}, w_{t-2}, w_{t-1} which contribute to the predictors of a "pentagram" ($n = 5$) language model with non-contiguous bigram and trigram estimators.

Each one of the 3×5 predictor sets described above has been tested both with linear and with rational interpolation. Whilst setting up one single interpolation coefficient per predictor was sufficient in the latter case, the linear interpolation scheme requires a set of separate coefficient vectors (see [8]) in order to deal with the problem of pathological ML estimators $\hat{P}_i$ occurring as soon as the frequency count in the denominator is vanishing.

4.1. Test Set Perplexities

Running perplexity evaluations on different text corpora, we found that rational language models consistently outperform linear ones. We present test set perplexities for trials with the INTERCITY train timetable inquiry corpus (Fig. 3) and the VERBMOBIL face-to-face business appointment dialog corpus (Fig. 4). Each of the two text corpora had previously been partitioned into three independent subsets: the *training* data to obtain the relevant n-gram counts, the *cross-validation* data to optimize the interpolation coefficients, and the *test* data to compute perplexities as a measure of the model's capability to generalize from the training data. The sizes of training/validation/test sets were 12921/900/2081 words for the INTERCITY inquiries, or 113321/1279/9009 words for the VERBMOBIL dialogs, resp.

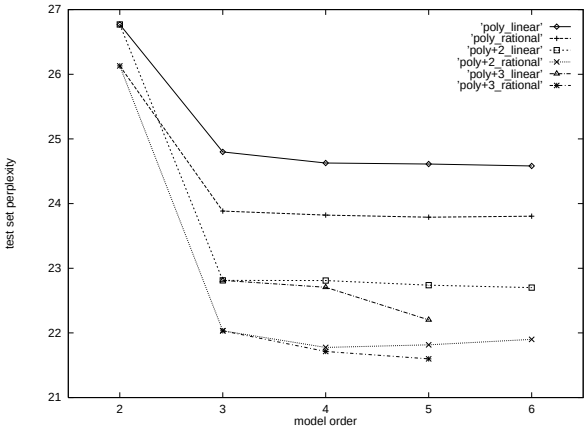

Figure 3. INTERCITY corpus

For both domains a substantial reduction in test set perplexity was achieved: The perplexity of the INTERCITY data was reduced by 12.2% (PPX 24.6 → 21.6), where a 4% improvement was due to rational interpolation and about 8% was due to the inclusion of non-contiguous bigram and trigram predictors. The overall perplexity reduction of the VERBMOBIL data amounted to 10.4% (PPX 103.7 → 92.9) where rational interpolation accounted for about 6% and non-contiguous predictors for the remaining 5%. Note that the extension of the predictor set strides along with a considerable increase of storage requirements for the "sparse" bigram and trigram statistics. The improvement caused by rational interpolation, however, was achieved without additional storage or computational cost compared with the linear one.

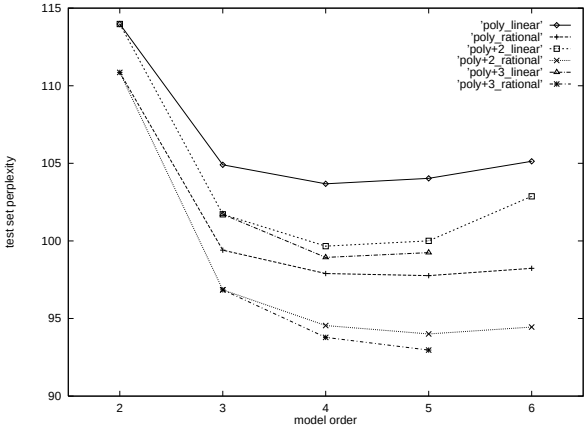

Figure 4. VERBMOBIL corpus

5. EXPERIMENTAL RESULTS II

Although test set perplexity is widely used as a measure of comparison for language models, perplexity reductions do not always result in improved recognition rates. In this section we will present experimental results that allow for a comparison of different interpolation strategies in three different recognition and classification tasks:

- spontaneous speech recognition
- dialog act classification based on subword units
- language identification based on framewise labeling

5.1. Speech Recognition

The most common goal of improving language model perplexity is the reduction of word error rates in automatic speech recognition. We performed experiments on the german spontaneous VERBMOBIL face-to-face business appointment dialog corpus, using a test set of 268 utterances with a total of 5065 reference words. We used a two–pass HMM word recognizer with a vocabulary size of 3833 words, incorporating bigram information in the first pass and polygram information in the

second pass [4]. The baseline error rate using linearly interpolated polygram models with $n = 3$ is 22.9%. Using rationally interpolated polygram models with hyperbolic weight functions (Section 3) and $n = 3$ results in an error rate of 22.3%. This is an error rate reduction of 2.4%. Although this improvement may appear rather small, it is important to note that is is achieved with practically no computational overhead, neither during language model training nor during decoding.

5.2. Dialog Act Classification

Although mostly used for speech recognition, n-gram models can be used for a wide range of tasks. One interesting application is the classification of dialog acts based on the recognizer output [9]. This allows for "flat analysis" of the user utterances, which is useful for keeping track of the dialog even when the linguistic analysis fails [1]. The idea is to segment user utterances into dialog acts using prosodic information. The corresponding parts of the recognized word sequence are then classified according to a ML–decision using different n-gram models for a predefined set of dialog acts. We performed experiments on a manually segmented and transcribed subset of the VERBMOBIL corpus. We used a set of 18 dialog act classes; the test set consisted of a total of 521 dialog acts. Using linearly interpolated polygram models, we achieved the lowest error rate of 36.4% with a polygram order of $n = 4$. A rationally interpolated polygram of order $n = 6$ augmented with time warped bigrams produced an error rate of 34.2%. This is an error rate reduction of 6%. The larger improvement compared to the word recognition experiments is due to the fact that dialog act classification is performed *only* based on language model information, while word recognition is based on language model information *and* acoustic information.

5.3. Language Identification

Another application for n-gram models we are investigating is language identification. As we cannot always assume to have a word recognizer or a phone recognizer for all languages of interest, we use a vector quantizer to assign one of 96 labels to each time frame of 30ms. The sequence of frames is then classified according to a ML–decision using n-gram models for each language [12]. For these experiments we used a set of 13 languages and speech signals of rather poor quality. The signals in the test sample are cut into fragments of only 2 seconds. On these fragments, we achieved a recognition rate of 37.3% using linearly interpolated polygram models of order $n = 3$; rationally interpolated polygram models augmented with time warped bigrams and $n = 3$ produced a recognition rate of 37.9%. This rather poor improvement is due to the fact that the language identification is mostly based on *counts* of specific symbols; the *sequence* of symbols contributes much less to the discrimination of different languages. This is confirmed by a recognition rate of 29% that can be achieved with $n = 1$ (unigram) models. Besides, the training data consist of a comparatively large amount of almost 500000 labeled time frames with only 96 different labels, which makes interpolation less important than with sparse training data.

6. CONCLUSION

In our paper, we addressed the problem of estimating stochastic language models based on n-gram statistics. We presented a novel approach, *rational interpolation*, for the combination of a competing set of conditional n-gram word probability predictors, which consistently outperforms the traditional linear interpolation scheme. The superiority of rational interpolation is substantiated by experimental results from language modeling (10–12% perplexity reduction), speech recognition (2.4% decrease in word error rate), dialog act classification (6% decrease), and language identification (1.6% improvement).

The training algorithm and the probability calculations of rational interpolation produce only neglectable additional cost in storage and computation when compared with the linear scheme. Moreover, there is good reason to suppose that our approach constitutes a promising alternative to the EM-based deleted interpolation algorithm for the tasks of smoothing stochastic decision trees and of interpolating context-dependent phone HMMs in speech recognizers.

REFERENCES

[1] H.U. Block. The Language Components in Verbmobil. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 79 – 82, München, 1997.

[2] K.W. Church and W.A. Gale. A Comparison of the Enhanced Good-Turing and Deleted Estimation Methods for Estimating Probabilities of English Bigrams. *Computer Speech & Language*, 5(1):19–54, 1991.

[3] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Royal Statist. Soc. Ser. B*, 39(1):1–22, 1977.

[4] F. Gallwitz, E.G. Schukat-Talamazzini, and H. Niemann. Integrating Large Context Language Models into a Real Time Word Recognizer. In N. Pavesic, H. Niemann, S. Kovacic, and F. Mihelic, editors, *Speech and Image Understanding*, pages 105–114. IEEE Slovenia Section, Ljubljana, Slovenia, April 1996.

[5] F. Jelinek. Self-Organized Language Modeling for Speech Recognition. In A. Waibel and K.F. Lee, editors, *Readings in Speech Recognition*, pages 450–506. Morgan Kaufmann, San Mateo, CA, 1990.

[6] F. Jelinek and R.L. Mercer. Interpolated Estimation of Markov Source Parameters from Sparse Data. In E.S. Gelsema and L.N. Kanal, editors, *Pattern Recognition in Practice*, pages 381–397. North Holland, 1980.

[7] S.M. Katz. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 35(3):400–401, 1987.

[8] T. Kuhn, H. Niemann, and E.G. Schukat-Talamazzini. Ergodic Hidden Markov Models and Polygrams for Language Modeling. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 357–360, Adelaide, Australia, 1994.

[9] M. Mast, R. Kompe, S. Harbeck, A. Kießling, H. Niemann, E. Nöth, E.G. Schukat-Talamazzini, and V. Warnke. Dialog act classification with the help of prosody. In *Int. Conf. on Spoken Language Processing*, volume 3, pages 1728–1731, Philadelphia, 1996.

[10] H. Ney and U. Essen. On Smoothing Techniques for Bigram-Based Natural Language Modelling. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 825–828, Toronto, 1991.

[11] H. Ney, U. Essen, and R. Kneser. On Structuring Probabilistic Dependences on Stochastic Language Modelling. *Computer Speech & Language*, 8(1):1–38, 1994.

[12] E. Nöth, S. Harbeck, and H. Niemann. Language Identification in the Context of a Human Machine D ialog System. In R. Kuhn, editor, *Proc. of the 3rd CRIM / FORWISS Workshop*, pages 85–95. Montreal, October 1996.

[13] P. O'Boyle, M. Owens, and F.J. Smith. A Weighted Average n-Gram Model of Natural Language. *Computer Speech & Language*, 8(8):337–349, 1994.

[14] G. Riccardi, R. Pieraccini, and E. Bocchieri. Stochastic Automata for Language Modeling. *Computer Speech & Language*, 10(4):265–293, 1996.

[15] E.G. Schukat-Talamazzini. Stochastic Language Models. In *Electrotechnical and Computer Science Conference*, Portorož, Slovenia, 1995.

[16] E.G. Schukat-Talamazzini, T. Kuhn, and H. Niemann. Speech Recognition for Spoken Dialog Systems. In H. Niemann, R. De Mori, and G. Hanrieder, editors, *Progress and Prospects of Speech Research and Technology*, number 1 in Proceedings in Artificial Intelligence, pages 110–120. Infix, 1994.