To appear in:
International Conference on Acoustics and Signal Processing (ICASSP '99),
Phoenix, Arizona, USA, March 1999

HMM AND NEURAL NETWORK BASED SPEECH ACT DETECTION

Klaus Ries

Interactive Systems Labs
Carnegie Mellon University, Language Technology Institute, Pittsburgh PA, 15213, USA
Universität Karlsruhe, Fakultät für Informatik, Lehrstuhl Prof. Waibel, 76128 Karlsruhe, Germany

ABSTRACT

We present an incremental lattice generation approach to speech act detection for spontaneous and overlapping speech in telephone conversations (CallHome Spanish). At each stage of the process it is therefore possible to use different models after the initial HMM models have generated a reasonable set of hypothesis. These lattices can be processed further by more complex models.

This study shows how neural networks can be used very effectively in the classification of speech acts. We find that speech acts can be classified better using the neural net based approach than using the more classical ngram backoff model approach. The best resulting neural network operates only on unigrams and the integration of the ngram backoff model as a prior to the model reduces the performance of the model. The neural network can therefore more likely be robust against errors from an LVCSR system and can potentially be trained from a smaller database.

1. INTRODUCTION

Speech act classification and a number of related problems that make use of a "chunk and label" paradigm have been studied by various authors in the recent past [NM94, WKNN97, REKK96, TKI+97, JBC+97a, FLL+98]. The basic idea is to use an HMM where the states are speech acts and the symbols emitted are sentences. While this study does not doubt that the underlying structure of the discourse model is an HMM but it questions how the output distributions of the sentences should be modeled (Fig. 1).

Currently all approaches to calculate the output distributions are based on ngram modeling: For each HMM state or speech act type an ngram backoff model is calculated. This requires the model to describe the full sentence distribution and irrelevant features may be harmful.

In contrast we suggest to do a direct classification of each speech act using neural networks. Neural networks have been succesfully tested on other related tasks such as parsing [BW96, Buø96, WW97] and classification of speech acts from prosodic events [SBC+98]. In this context we will see that the networks can be interpreted in a fashion that would make them a natural extension to standard ngram modeling. A hybrid approach using both neural networks and ngram backoff modeling is introduced and tested. It will be shown how they can be effectively integrated into a full HMM speech act classification system and how our current system is operating.

This research has been funded in part by a grant from the US department of defense. My foremost thanks go to my advisors Alex Waibel, Alon Lavie and Lori Levin. I am grateful to use the JANUS-Rtk for my experiments, especially the neural network component by Jürgen Fritsch. Klaus Zechner has built the Spanish POS tagger and we are using Eric Brills POS software. Our discourse taggers, Liza Valle, Santiago Cortes, and Susana Fuentes Arce have been working really hard to make everything happen. I would also like to thank everybody at the Interactive Systems Labs and the LVCSR summer workshop 97 for the discussion, support and fun I have found there.

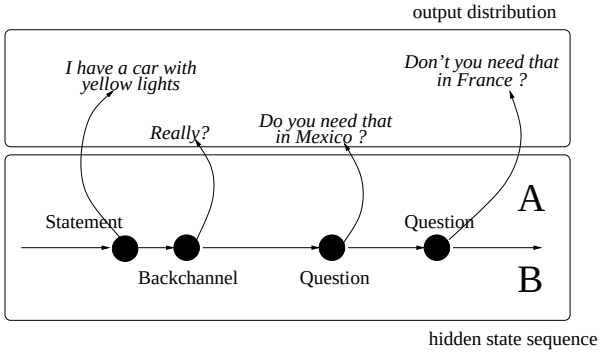

Figure 1: **HMM of speech acts:** The underlying structure of the discourse model is a hidden Markov model. The output distributions are distributions over sentences rather than words. This paper focuses on neural network based estimates for the output distributions.

2. HMM SPEECH ACT CLASSIFICATION

The basic idea of an HMM based speech act classification system is to model the speech acts as hidden events that need to be identified using a Viterbi search through the HMM. The HMM states are emitting the words in a speech act at once and – unless the segmentation into the speech acts is given – the speech act boundaries between words also have to be found [NM94, WKNN97, REKK96, TKI+97, JBC+97a, FLL+98]. Since speaker can overlap, produce more than one speech act between two pauses and can produce speech acts that last over a pause the segmentation problem is not trivial in spontaneous speech databases. One of the major problems of this speech act classification approach is that the full sentence distribution needs to be modeled for all speech acts and this problem will be addressed in this paper.

In the first step we have been looking only at one channel at a time. As we have reported in [FLL+98] that the problem of segmentation does not have to be solved in conjunction with the prob-

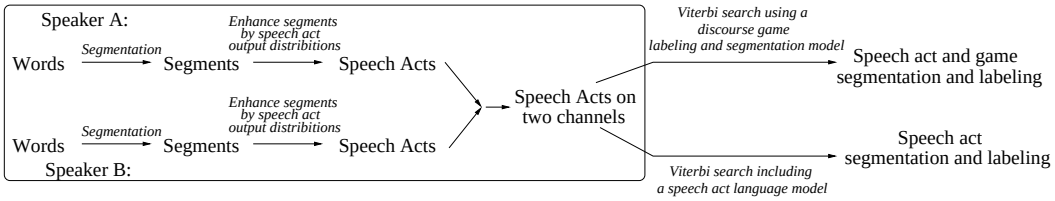

Figure 2: **Incremental lattice construction:** The input to the discourse model can be a lattice of words, e.g. produced by a speech recognizer. A lattice of segments is produced by an A^* search over the input lattice that can hypothesis segment boundaries and inserts every expanded segment in the segment lattice. The speech act lattice is generated by replacing each segment by the set of all possible speech acts and assigning a likelihood for each segment given the speech act. The final speech act sequence is found by a A^* Viterbi search operating over the lattices from both channels. The A^* search is also capable of segmenting and switching language models according to hidden states, e.g. a dialogue game model operating on speech acts.

lem of labeling the speech acts. We have therefore been able to produce equally good segmentation and labeling results by producing a segment lattice [1] and restricting our search: The segment lattice is currently produced by an A^* search procedure that can hypothesize segment boundaries. We have optimized the lookahead function to reflect the future effects of these segment boundaries. We produce an entry in the segment lattice if the A^* procedure is completing one segment. We have also used the same A^* procedure to search over the word lattice while simultaneously segmenting and labeling it with speech acts according to the speech act HMM with ngram models as the sentence distribution. The results are the same as if we restrict our search to the lattice of speech acts generated from the segment lattice.

In a second step we looked at two channels with possible speaker overlap. In our domain, CallHome Spanish, overlap is extremely frequent and is often long. Since it has been reported [JBC$^+$97b] that ngrams of speechact/speaker information ordered by time is constituting a good model of speech act sequences we constructed a search that could search over segmentations and labels on both channels. The procedure therefore takes a segment graph for each channel as its input. The segment graph is expanded by replacing each segment by a segment/speechact pair with the correct emission probability for the sentence given the speech act. The search can then proceed over these two lattices in parallel and apply a language model on the speech acts in the search. The search is restricted such that beginning times of the speech acts, even if they belong to different channels, are ascending (Fig. 2). The application of the language model that can operate over multiple speakers did not yield the performance gain we hoped for, although preliminary experiment indicated a reduction in perplexity and others have reported improvements [JBC$^+$97b].

In principle it is straightforward to extend this approach to higher level discourse processing such as dialogue game classification. Indeed our current implementation of the A^* is able to search over a lattice and it can switch between speech act specific language models according to a hidden state and at the same time enforce a standard ngram model on the hidden states. This would allow us to do speech act labeling and segmentation directly from the lattice but it also allows us to do speech act and dialogue game [LTGR$^+$98] classification and segmentation from the two speaker speech act lattice. Since the A^* procedure also allows to generate a lattice of dialogue games this process could be iterated to higher levels. The dialogue game lattice would have the practical advantage that it is a single lattice operating on one time scale over one channel as opposed to the speech act or segment lattice.

The next step we want to take is the integration of prosodic information into the classification process. This can be done by using alternative output distributions in the expansion of the segment lattice to the speech act lattice (Fig. 2). Earlier [FLL$^+$98] we already suggested to apply a hybrid model which takes into account the information of the backoff models as a prior and tries to improve the performance of the classifier directly. This idea was born due to two major observations: The decoupled integration of ngram backoff model information with prosodic information did not yield strong effects on the classification performance [JBC$^+$97b] and the insight that ngram backoff models do model the most important prosodic feature – the length of the utterance – fairly well [FLL$^+$98]. Additional we preferred the use of neural networks since they would allow us to interpret their outputs as probabilities and that they have given us very competitive results on prosodic speech act classification previously [2]. The neural network we have been using is a three layer network with "shortcut"-connections from the input the output layer and a small number of hidden units. The output layer is using the softmax function [Jor95] while the activation function of the hidden layer is $\tanh$, the error function is the Kullback-Leibler divergence. Each output unit could receive an additional input that could be determined by another knowledge source. This can be interpreted as a prior on the output distribution: It allows to build a hybrid model that takes into account the probability distribution of ngram backoff models in the neural network. Formally the model is

$$y(x,z) = exp(A \cdot x + B \cdot \tanh(C \cdot x)) \cdot \hat{y}(z)/Z(x,z)$$

The A is the weight matrix for the shortcut connections, B and C for the hidden layer, $\hat{y}(z)$ is a prior and $Z(x,z)$ is chosen such that $\sum_x y(x,z) = 1$

Assuming that the input nodes consist of all unigrams the connection from a word w to each output speech act sa could be set

[1] As common in the speech recognition community a lattice is the term used for a directed acyclic graph with a start node that can reach all nodes in the lattice and an end node that can be reached from all nodes.

[2] These earlier experiments did not yet include the **RPROP** training algorithm that we have been using in this study and we would assume that the results could be improved with this technique

to $log(p(w, sa))$ and we would obtain exactly the classifier for a unigram model on the speech acts and the sentences conditioned by the speech act.

The output of a (neural net based) classifier $p(sa|W)$, where sa is the speech act and W is the set of words in it, can be rewritten using Bayes Rule to

$$p(W|sa) = \frac{p(sa|W) \cdot p(W)}{p(sa)}$$

and can therefore be used as the output distribution in an HMM system. The segment lattice allows the effective integration into an overall system that takes multiple segmentations into account. Other experiments on segmentation with neural networks have so far provided us with mixed results [FLL+98, GZA97] compared to the HMM based approach and are worth reviewing after the results of the experiments we report here.

3. EXPERIMENTS

The incremental lattice construction scheme has been tested with ngram backoff models as output distributions and it gave us the same results (if we searched on each lattice separately) as if we would search with a model with a hidden state (speech act) and segmented the lattice directly (compare e.g. [WKNN97, FLL+98]). Using discourse models over both channels did not yield improvements in detection accuracies so far in contrast to [JBC+97a]. Since the perplexity of discourse models taking into account information from both channels is much lower than that of a single channel model and the detection results are not worse we are keeping this approach.

The following neural network experiments we have done are only covering the simple speech act classification with a known segmentation and without a discourse model. The results of these classifiers can be integrated in the expansion of our segment graph and we therefore measure here to what extent the output distribution of the HMM can be optimized for performance. We also have not tested the combination of prosodic features with the word based information yet, however the use of similar networks (yet without the RPROP training algorithm used successfully here) has been shown to consistently slightly outperform decision tree based methods [SBC+98] and the prosodic features could be added in a straightforward fashion. Both the searches discussed as well as standard prosodic feature algorithms are integrated in the JANUS-Rtk [ZFRW97].

As reported in [FLL+98, EDR98] we are using a parts of speech (POS) tagger to annotate the database and we map all but the most frequent 250 word/tag pairs on their POS tag since these are the most important features for an ngram based system. To be able to use more than just unigrams we incorporated salient sequences [Gor95] in a simple greedy bottom up clustering procedure and labeled them in the database.

We have tested several variations of training algorithms and we have clearly received the best results using RPROP [RH93] including an extension equivalent of weight decay. Without RPROP we could get reasonable results using training by pattern algorithms and a carefully adjusted the learning rate. Using RPROP without weight decay we ran into overtraining if we did not stop early enough but the error function decrease significantly over the standard backpropagation algorithm. The weight decay for RPROP practically eliminated the overtraining problem. The implementation of this variant of RPROP follows the one featured in SNNS [Zel93] and has been added to the neural network library in JANUS-Rtk [ZFRW97].

All models were trained on 55 dialogues and tested on 40 dialogues, corresponding to 25500 respectively 14400 speech acts from the CallHome Spanish database. Neural networks with no salient words (basically just unigram features) alone result in significantly better results than the the ngram backoff models. Adding ngram backoff models in the prior distribution has hurt the performance in all cases we investigated. We also observed that the backoff models are prone to overtraining: The trigram model is decreasing in performance and we also observed that, unlike the neural networks, the performance of the bigram and trigram backoff model is much better on the training than on the test set (see Table 1).

Classifier	classification accuracy
baselines	
pick the most likely speechact	40.0%
NN, shortcuts, 3 hidden units, length of speechact	48.8%
ngram models	
unigram model	72.9%
bigram backoff model	74.4%
trigram backoff model	70.0%
neural networks	
shortcuts, no hidden units, unigram features	75.4%
shortcuts, 3 hidden units, unigram features	76.2%
shortcuts, 5 hidden units, unigram features	75.9%
no shortcuts, 3 hidden units, unigram features	73.8%
shortcuts, 3 hidden units, unigram+salient sequence features	75.6%
hybrid neural networks with shortcuts and 3 hidden units	
unigram features, unigram prior	76.1%
unigram features, bigram prior	74.8%

Table 1: **Speech act classification results:** A simple neural network outperforms the ngram backoff model, adding more features or priors to the neural network does not increase the performance

4. CONCLUSION

The original idea of these experiments was to efficiently integrate prosodic features with word based features. Much to our surprise we have found not only that the neural network alone is as good as the ngram backoff model classifier but it outperforms it. This is even more astonishing if we compare this to the rate of improvement we have seen from other techniques such as context modeling [FLL+98, JBC+97b] and prosodic feature integration [SBC+98] and keep in mind that the highest possible accuracy is limited by the intercoder agreement [3].

[3] We have measured an intercoder agreement that is not exactly the one we would need for the comparison. By rough comparative measurements with SWBD annotations we assume that it is similar or worse than the one for SWBD. The intercoder agreement on SWBD was 83% so one cannot assume that any system could surpass this mark since it also represents the annotation error.

Additionally the features we have been using are simple features (unigrams of frequent words and POS tags) that we would assume to be detected more reliably by an LVCSR system.e Neural network based parsing algorithms (see e.g. [BW96, Buø96, WW97] and its references) made different choices in the feature representation and this surprisingly good results sheds a new positive light onto the application of neural networks in parsing and speech understanding. This holds even more since models similar to our speech act detection model have been used by many authors (e.g. [Min97]) to do speech understanding. The HMM backbone, incremental lattice generation component, learning algorithms and feature representation of this hybrid system might on the other hand contribute to the neural network parsing community.

In the nearby future we will have integrated our complete system and have done experiments with prosodic features. The completion of the discourse game tagging effort will enable us to run discourse game detection over the speech act lattice. Additional we have to revisit the input representation of our classification model and the learning techniques applied since neural network based parsing systems have made different choices that might prove helpful here as well [BW96, Buø96, WW97].

5. REFERENCES

[Buø96] Finn Dag Buø. *FeasPar - A Feature Structure PARser learning to parse spontaneous speech*. PhD thesis, University of Karlsruhe, 1996.

[BW96] Finn Dag Buø and Alex Waibel. Feaspar : A feature structure parser learning to parse spoken language. In *COLING*, 1996.

[EDR98] Carol Van Ess-Dykema and Klaus Ries. Linguistically engineered tools for speech recognition error analysis. In *ICSLP*, 1998.

[FLL+98] Michael Finke, Maria Lapata, Alon Lavie, Lori Levin, Laura Mayfield Tomokiyo, Thomas Polzin, Klaus Ries, Alex Waibel, and Klaus Zechner. Clarity: Automatic discourse and dialogue analysis for a speech and natural language processing system. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, March 1998.

[Gor95] Allen Gorin. On automated language acquisition. *Journal of the Acoustical Society of America*, 97(6):3441–3461, June 1995.

[GZA97] Marsal Gavalda, Klaus Zechner, and Gregory Aist. High performance segmentation of spontaneous speech using part of speech and trigger word information. In *Fifth Conference on Applied Natural Language Processing*, Washington,DC, 1997.

[JBC+97a] Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. Automatic detection of discourse structure for speech recognition and understanding. In *IEEE Workshop on Speech Recognition and Understanding*, September 1997.

[JBC+97b] Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. SWBD Discourse Language Modeling Project, Final Report. Technical report, Johns Hopkins LVCSR Workshop-97, 1997.

[Jor95] Michael Jordan. Why the logistic function? a tutorial discussion on probabilities and neural networks. `ftp://psyche.mit.edu/pub/jordan/uai.ps.Z`, 1995.

[LTGR+98] Lori Levin, Ann Thyme-Gobbel, Klaus Ries, Alon Lavie, and Monika Woszczyna. A discourse coding scheme for conversational spanish. In *ICSLP*, 1998.

[Min97] W. Minker. Stochastically-based natural language understanding across tasks and languages. In *Eurospeech*, 1997.

[NM94] M. Nagata and T. Morimoto. First steps towards statistical modeling of dialogue to predict the speech act type of the next utterance. *Speech Communication*, 15:193–203, 1994.

[REKK96] N. Reithinger, R. Engel, M. Kipp, and M. Klesen. Predicting dialogue acts for a speech-to-speech translation system. In *ICSLP*, 1996.

[RH93] M. Riedmiller and Braun H. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *Proc. of the IEEE Int. Conf. on Neural Networks*, pages 586–591, 1993.

[SBC+98] Elizabeth Shriberg, Rebecca Bates, Noah Coccaro, Daniel Jurafsky, Rachel Martin, Marie Meteer, Klaus Ries, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, forthcoming, 1998.

[TKI+97] Paul Taylor, Simon King, Stephen Isard, Helen Wright, and Jacqueline Kowtko. Using intonation to constrain language models in speech recognition. In *EUROSPEECH*, Rhodes, Greece, 1997.

[WKNN97] V. Warnke, R. Kompe, H. Niemann, and E. Nöth. Integrated dialog act segmentation and classification using prosodic features and language models. In *Eurospeech*, pages 207–210, 1997.

[WW97] Stefan Wermter and Volker Weber. Screen: Learning a flat syntactic and semantic spoken language analysis using artificial neural networks. *JAIR*, 6:35–85, 1997.
http://www.jair.org/abstracts/wermter97a.html.

[Zel93] Andreas Zell. Snns user manual, version 3.0. Technical Report 3, University of Stuttgart, Institute for Parallel and Distributed High Perfomance Systems, 1993.

[ZFRW97] Torsten Zeppefeld, Michael Finke, Klaus Ries, and Alex Waibel. Recognition of conversational telephone speech using the janus speech engine. In *Proceedings of the ICASSP'97*, München, Germany, 1997.