In: *4th European Conference on Speech Communication and Technology (EUROSPEECH '95), Madrid, Spain, 18 - 21 September 1995*

INTEGRATING SPELLING INTO SPOKEN DIALOGUE RECOGNITION

Hermann Hild and Alex Waibel

Interactive Systems Laboratories
University of Karlsruhe — 76128 Karlsruhe, Germany
Carnegie Mellon University — Pittsburgh, USA

ABSTRACT

Recognition of spelled letter sequences is essential for many real-world applications which involve arbitrary names or addresses. Often the letter sequences carry the sentence's crucial information; therefore, it is important to correctly localize and recognize the spelled string. However, large vocabulary speech recognizers tend to perform poorly on spelled letters, especially if they have to deal with spontaneous speech. The research presented here aims at improving the recognition accuracy of spontaneous speech with embedded spelled-letter sequences. We propose methods to localize spelled-letter segments and reclassify them with a specialized letter recognizer.

1. INTRODUCTION

Applications of spelling in spontaneous speech include the recognition of spelled names or addresses, as well as *repair dialogues*, where spelling can be used to disambiguate the inevitable recognition errors or to introduce out-of-vocabulary words to the system [4]. All these applications have in common that in natural speech, sentences carrying spellings will rarely consist of plain, clean letter sequences. Rather, the spelling is embedded in spontaneous speech, which introduces, corrects, comments on, paraphrases or further specifies the individual letters. Since it is crucial to reliably extract the letter sequence from the sentence, it is important to develop robust recognition procedures for spelling dialogues. In first experiments towards this task, we use the following approach. First, the entire sentence is recognized with a large vocabulary spontaneous speech recognizer. Second, potential letter segments are identified. These segments are then reclassified with a specialized letter recognizer.

2. THE SPONTANEOUS SPEECH RECOGNIZER

The recognition front-end of JANUS [6], the speech-to-speech translation system of Carnegie Mellon and Karlsruhe University was used for the experiments described below. JANUS has a large vocabulary spontaneous speech recognizer engine with built-in inter-word noise models for more robust acoustic modeling of spontaneous speech.

3. THE LETTER RECOGNIZER

A connectionist recognizer, the Multi-State Time Delay Neural Network (MS-TDNN) [2, 3] is used for connected spelled letter recognition. The MS-TDNN integrates the time-shift invariant architecture of a TDNN [5] and a nonlinear time alignment procedure (DTW) into a word-level classifier. The front-end TDNN uses sliding windows with time-delayed connections to compute a score for each phoneme-like state in every frame. Each word to be recognized is modeled by a sequence of phonemes; in a dynamic time warping procedure, an optimal alignment path is found for each word. The activations along these paths are then collected in the word output units. The error derivatives are backpropagated from the word units through the alignment path and the front-end TDNN. For continuous recognition, the standard "one stage dynamic programming search" is used. Several other search techniques can improve recognition performance if a closed list of spelled words is given [1]; however, we are not using such strong constraints for the experiments reported below.

The MS-TDNN was trained and crossvalidated on a data base of 9128 German strings (57301 letters) spelled by 80 speakers. On a test set of 2213 strings (15035 letters) spelled by another 20 speakers, a letter recognition rate of 90.1% was achieved (without using any language model).

4. EXPERIMENTAL SETUP

The VERBMOBIL Evaluation Test Set from June 1994 comprises 842 German sentences from various spontaneous scheduling task dialogues recorded at four different sites. 76 of the 842 sentences contain spellings. The 2533 words of those 76 sentences in-

clude a total of 653 letters in 111 segments. Basically, in each of the 76 sentences one last name is spelled. There are more letter segments than sentences because some spellings are separated by non-letter words such as "Strich" (hyphen), "doppel" (double), "scharf" (for the German "scharfes S" (ß)) or other filler words like "and then". Also, some people spell "A wie Anton, B wie Berta ..." ("A as in ...").

In addition to the regular letters, the letter recognizer contains some of the most typical "between-letter" words in its vocabulary <LET>:

```
<LET> =
 { A,...,Z                          ; Letters
   Ä, Ö, Ü,                         ; "Umlaute"
   Eszet, scharf S, scharfes S,     ; German ß
   Strich, Bindestrich              ; Hyphen
   doppel }                         ; Double
```

If all words in <LET> are counted as letters, the number of letters increases to 668 and the number of letter segments reduces to 92.

5. BASELINE RESULTS

The current performance results of the JANUS system are 68.6% and 71.4% word accuracy, achieved on the VERBMOBIL June 1995 evaluation sets of 66 "long" and 256 "short" sentences, respectively.

The baseline system used for these experiments is an earlier JANUS System as used for the 1994 VERBMOBIL evaluation[1], which achieved 45% word accuracy on the 842 sentences of the June 1994 VERBMOBIL evaluation set. The large performance gap between the two systems is due to a more difficult 1994 test set (over 12% out-of-vocabulary words), a smaller training set and an earlier version of the JANUS system.

6. LOCALIZING LETTER BOUNDARIES

Most spellings contain more than one letter. To get more robust letter recognition, this fact can be exploited by forcing JANUS to recognize at least three letters in a row. To enforce this constraint, the language model was modified so that once the search finds a letter, it cannot leave that path before at least three letters in a row were recognized. This constraint was incorporated into the language model, as indicated in figure 1.

[1] As the new JANUS was only recently available and new spelling data were only recently released, no results using the new JANUS system could be incorporated in this paper. However, we will report these results in a later publication.

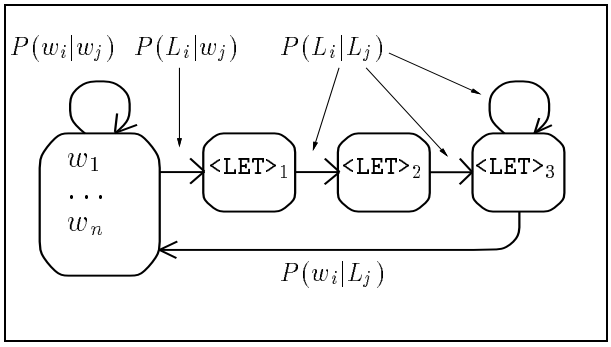

Figure 1: Language model to force recognition of at least 3 letters in a row.

Transitions within non-letter words $\{w_1, \ldots, w_n\}$ are the bigrams of the JANUS language model. Transitions within the 35 letters <LET> $= \{L_1, \ldots, L_{35}\}$ are computed as bigrams from a list of 110,000 (43000 unique) last names or are assumed to be equally probable. Since there is only a small amount of data to estimate the language model statistics for spell sentences, transitions from regular non-letter words w_i into letters L_j were generalized to transitions into the letter class <LET>:

$$p(L_j|w_i) := p(\texttt{<LET>}|w_i) * p(L_j|\texttt{<LET>})$$

The above definition is correct only if the words used to introduce a spelling and the first letter of the spelling are independent, which is a reasonable assumption[2].

With the baseline JANUS system tested on the 76 sentences with spellings, a word accuracy of 39.8% (measured on the entire sentence) and a "letters-only" accuracy (i.e. only recognized letter segments are matched with the correct spellings) of 22.7% was achieved. Besides many letter-letter confusions, the letters are easily inserted or confused with other words in the sentences. Using the "3 forced letter technique" improved the recognition of JANUS on the 76 sentences to 45.9% word accuracy and to 27.2% letters-only accuracy.

When the found letter segments were reclassified with the letter recognizer, a word accuracy of 51.8% and a letters-only accuracy of 47.0% was achieved. Of course the boundaries of the found letter segments are not exact, they may miss letters or include non-letter words at their beginning and end. If the "letters-only" accuracy is measured within the correct letter segment boundaries, 77% letter accuracy is achieved. With the Karlsruhe train and test spell

[2] Sometimes a last name is first spoken and then spelled. In these cases the last word before the spelling is correlated to the spelling. However, arbitrary last names are out-of-vocabulary words and therefore cannot (yet) be utilized to exploit this dependency.

System	Overall WA	Letters-only
JANUS	39.8 %	22.7 %
JANUS + 3 forced letters	45.9 %	27.2 %
JANUS + 3 forced letters + spelling reclassification	51.8 %	47.0 %
JANUS + correct letters enforced	57.6 %	92.0 %

Table 1: Word Accuracy on the 76 spelling sentences with various system configurations, measured on all words in the sentences and on the letter sequences only.

data, the letter recognizer reaches about 90% letter accuracy. The loss in performance in the VERBMOBIL test set can be explained by the fact that the spellings are more noisy and spontaneous; also they were recorded at different sites, while the letter recognizer was trained site dependent.

The recognition errors in spelled-letter regions cause subsequent errors. To determine an upper bound on how much can be gained by perfect letter recognition, we designed the following experiment: For each sentence the language model is modified in such a way that if the recognizer enters a letter sequence, only the spelling present in this particular sentence can be recognized. However, if any spelling is recognized, it can occur once or several times at any position in the sentence. Therefore, a close to perfect letter recognition is simulated. Interestingly, in two cases, an out-of-vocabulary proper name ("Gerholzner" and "Zelthofer") was recognized as the corresponding spelling (G E R H O ...) of that proper name. Also, eight sequences of letters (most of them only one or two letters long) were not found, resulting in a "letters-only" accuracy of 92.0% and a word accuracy of 57.6%. However, the frequent out-of-vocabulary proper names are still another significant source of errors in the spelling sentences.

7. CONCLUSIONS

Recognition of spelling embedded in spontaneous speech is a challenging task of practical importance for many applications. We have presented methods towards a more robust recognition of spelling dialogues. The results of our first experiments are summarized in Table 1. We could improve recognition of spelling sentences from 40% to 52%. Unfortunately we had to use a relatively early JANUS version as a baseline system. Results of experiments using the June 95 baseline system will be reported in a later publication.

Acknowledgements

This research was partly funded by SIEMENS AG and by grant 413-4001-01IV101S3 from the German Ministry of Science and Technology (BMFT) as a part of the VERBMOBIL project. The authors would like to thank all members of the Interactive Systems Labs, especially Michael Finke, Petra Geutner, Thomas Kemp, and Monika Woszczyna, for discussions and assistance with the JANUS System.

8. REFERENCES

[1] Martin Betz and Hermann Hild. Language Models for a Spelled Letter Recognizer. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, May 9-12 1995.

[2] P. Haffner, M. Franzini, and A. Waibel. Integrating Time Alignment and Neural Networks for High Performance Continuous Speech Recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*. IEEE, May 1991.

[3] Hermann Hild and Alex Waibel. Speaker-Independent Connected Letter Recognition With a Multi-State Time Delay Neural Network. In *3rd European Conference on Speech, Communication and Technology (EUROSPEECH) 93*, pages 1481 – 1484, September 1993.

[4] Arthur E. McNair and Alex Waibel. Improving Recognizer Acceptance Through Robust, Natural Speech Repair. In *International Conference on Speech and Language Processing*, pages S22–15.1, 1994.

[5] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. Phoneme Recognition Using Time-Delay Neural Networks. *IEEE, Transactions on Acoustics, Speech and Signal Processing*, March 1989.

[6] M. Woszczyna, N. Aoki-Waibel, F.D. Buø, N. Coccaro, K. Horiguchi, T. Kemp, A. Lavie, A. McNair, T. Polzin, I. Rogina, C.P. Rose, T. Schultz, B. Suhm, M. Tomita, and A. Waibel. Janus 93: Towards Spontaneous Speech Translation. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 345–349, May 1994.