RECOGNITION OF SPOKEN AND SPELLED PROPER NAMES

Michael Meyer and Hermann Hild

Interactive Systems Laboratories
University of Karlsruhe — 76128 Karlsruhe, Germany
{hhild,mmeyer}@ira.uka.de

ABSTRACT

Many speech applications, most prominently telephone directory assistance, require the recognition of proper names. However, the recognition of increasingly large sets of spoken names is difficult: Besides technical limitations, very large recognition vocabularies contain many easily confused words or even homophones. Therefore, proper names are often spelled or both spoken and spelled.

In this paper we compare the performance for proper name recognition when a name is spoken only, spelled only, or both spoken and spelled. In the latter case, information about the same name is provided in two different representations. We address methods to exploit this redundancy and propose techniques to handle the recognition of large lists of spoken and spelled proper names.

1. SPOKEN AND SPELLED NAMES

1.1. Scenarios

In what contexts do people speak and spell names? Three scenarios of increasing complexity can be distinguished, as exemplified in table 1. In the most simple case, the spoken and spelled name are two separately recorded utterances (scenario 1). In a more user-friendly dialogue, one may be allowed to speak and spell "in one piece", i.e. without necessarily pausing in the same recording (scenario 2). Finally, the most challenging situation arises when the spelled and spoken name is embedded in spontaneous speech (scenario 3). In our experiments we will examine the first two scenarios.

1.2. Speech Data

We have collected a database of about 2800 German last names (randomly selected from a telephone directory of 100,000 names) spoken by 57 different speaker, according to scenario two: Each name was continuously spoken and spelled in one utterance, and recorded with a close-talking microphone at a sampling rate of 16 kHz. The speakers were instructed that no pauses are required between letters or between the spoken and spelled name.

To be able to conduct simulated scenario 1 experiments, the boundaries between all spoken and spelled names were identified. However, although each part can now be recognized separately, the situation is still somewhat different from real scenario 1 recordings, because there are potential coarticulations across the boundaries.

1.3. Pronunciation Dictionary

The recognition of proper names is an essential, but non-trivial problem. German telephone listings comprise about 30 million entries with about 1 million different names. Before each name can be added into the recognizers vocabulary, its pronunciation, i.e. its phonetic transcription must be known.

A subset of the ONOMASTICA database containing about 200,000 pronunciations of German last names was provided to us by courtesy of Deutsche Telekom and TU Berlin, covering about half of the 2800 names of our speech database. This set of 1337 spoken and spelled names is used for all the experiments described below.

1.4. The Human Factor

Compared to fluently speaking, spelling is a more robust but less natural mode to communicate a proper name. About 4% of the 2800 last names were spelled incorrectly, although the name to be spoken and spelled was presented in written form. A typical class of errors were omitted letters. Interestingly, this phenomenon often happens if the sound of the omitted letter is already present in the previous letter (e.g. 'E' after 'D', or 'A' after 'K') as for example observed in "`karolus k r o l u s`". Similarly, in "`campos k a m p o s`" or "`vogel f o g e l`", the erroneous spelling is oriented closer to the sound than to the actual orthography of the name.

1.5. Outline of Experiments

In the following sections, we will first describe the spelled letter and large vocabulary recognizers used for our experiments, as well as their results on each of the spelled and spoken part of the names by itself (section 2).

We will then discuss methods for combining the recognition of the spelled and spoken name. Depending on the size of the list of names to be recognized, different techniques are applied. If the list of names to be recognized is reasonably small, they can be kept in the recognizer's dictionary (section 3). However, with increasingly larger lists, we need to switch to methods which do not require to maintain all names in the dictionaries (section 4). Finally, we will describe some experiments where it is not a priori known if a name is spelled only, or spoken and spelled (section 5).

1.6. Related Work

Work on name retrieval from spellings is reported by many researchers. Cole et. al. [2] use individually scored letters to search names in a tree-structured database of

(1)	Please speak your name:	"**Smith**" Please spell it: "**S M I T H**"
(2)	Please speak and spell your name:	"**Smith, S M I T H**"
(3)	What's your name? :	"My name is **Smith**, that's **S M I T H**"

Table 1: Three scenarios for speaking and spelling a proper name

50,000 names. Junqua et. al. [6] employ a sophisticated multi-pass strategy to narrow down the list of name candidates. In [1] we compare several methods to constrain the search to a given list of names. Best results were achieved on a conceptionally simple tree-based method which is demonstrated on very large name lists in [5].

A comparison of spoken and spelled name recognition is presented by Kamm et. al. in [7]. Both spoken and spelled names are used for name retrieval in the telephone directory assistance system of Kaspar [8]. However, we are not aware of any literature which tries to explicitly combine the recognition of spoken and spelled names.

2. THE RECOGNIZERS

For our experiments, we use a Multi-State Time-Delay Neural Network (MS-TDNN) as a specialized letter recognizer, and the large vocabulary continuous speech recognition front-end of the JANUS Speech-to-Speech Translation System.

2.1. JANUS

The JANUS recognizer was trained and tested on the 1996 Verbmobil Evaluation data (a spontaneous scheduling task with a 5000 word vocabulary), achieving a word accuracy[1] of 86.2% in the official 1996 Verbmobil test set [3].

Using a pronunciation dictionary derived from the ONOMASTICA data, 60.0% names correct were achieved on the test set of the 1337 spoken last names. Compared to the spontaneous scheduling task, the loss of performance can be explained by several factors. First of all, there is no language model, resulting in a high perplexity of over 900, compared to about 50 for the Verbmobil task. The recognizer was never trained on isolated speech, but on continuous, spontaneous speech, which is quite a different speaking style. Also, it is unclear to what degree the pronunciations in the ONOMASTICA dictionary are consistent with the conventions used for the phonetic transcriptions of the JANUS dictionary. In addition, the recognizer was never trained on any of the words to be recognized, which may be especially a problem for the many non-German last names in the list (see table 2).

To recognize the spelled name with JANUS, each dictionary entry represents the phonetic transcription of the spelled name, e.g. "`[Lang E L - AH - E N - G EH]`" for the name "`Lang`". Given the list of 1337 names, 93.3% correct names were achieved on the spelled names. Obviously, spelled names can be much more robustly recognized than fluently spoken names.

Abel Abendschein Adams Adler Agha Akkoc Aksu Albiez Alesi Alexakis Alilovic Allgeier Alphan Ammersbach Anselm Apostolidis Appelt Artuso Asmus Attrasch Aubert Augustin Avci Aydogan Azad Böhm Böhme Böhnke Büchner Bühler Bürk Bacher Baier Baltz Baranowski Baron Barteczko Barthlott Bartholomaeus Bartl

Table 2: List of the first 30 of the 1337 last names in the test set

2.2. MS-TDNN

The MS-TDNN is an extension of the Time-Delay Neural Network. Similar to NN-HMM Hybrids, the MS-TDNN employs word[2] models and a dynamic time-alignment (DTW) to handle the time varying nature of the speech signal. However, in the MS-TDNN, the DTW is directly integrated in the connectionist architecture and training scheme, allowing for discriminant training on the word and sentence level [4]. Without using any language modeling, the speaker-independent recognition rate of continuously spelled letter sequences is about 90% letter accuracy.

We have experimented with various techniques to recognize spelled names from large lists of names [1]. The most successful approach turned out to be a search in which all spelled names are compiled into one large tree structure. With a time-synchronous search and no backpointers needed, a very efficient search can be implemented, allowing to recognize names from list sizes up to about 1 million names in real time [5].

Using this tree search approach, the MS-TDNN achieved 96.5% correct names on the 1337 spelled names from the test set.

3. SMALL LISTS

In this section we assume that the list of names to be recognized is small enough, so that every name can be explicitly represented in the dictionary, using the pronunciations provided by the ONOMASTICA dictionary.

How can we advantageously combine the different information provided by the spoken and spelled names? The two representations are not as orthogonal as one might think. After all, the pronunciations of the spelled letters represent in a first approximation the sounds of the letters in the fluently spoken words. For example, the acoustic realization of "`Tom`" versus "`T-O-M`" are quite

[1] For the sake of recognition speed, we were using a system with about 2% lower word accuracy.

[2] letters in our case

	Separate Recognition		**Combined** Recognition	
	fluently spoken (F)	spelled (L)	Scenario 1 (F+L)	Scenario 2 (FL)
1337 names in dictionary	60.0	96.5	97.7	95.8
multi-pass, 1337 names	-	96.5	97.7	96.9
multi-pass 100,000 names	-	87.1	89.5	88.1

Table 3: Summary of results for the separated and combined recognition of fluently spoken and spelled last names

similar. Exceptions are letters with "unusual" pronunciations, such as "double-U" or (in German) "Ypsilon", and those letter combination which define their own pronunciations, such as (in German) "`sch, ch, ck, th, pf, ph, ie`".

Capturing these relations in explicit rules is a quite difficult and probably not very promising strategy. In the following we will use a far less complex approach, which combines the two different representations on the basis of their acoustic scores only.

3.1. Scenario 1

We first consider the situation where we have two isolated utterances for the spoken and spelled name (scenario 1). Let $Y_L(i)$ be the score of a spelled name i found in the N-best list of the MS-TDNN letter recognizer, and $Y_F(i)$ the score of the same name in the N-best list of fluently spoken names as found by JANUS. The position of name i in the combined N-best list is determined by its new score

$$Y(i) = \lambda \cdot Y_L(i) + (1 - \lambda) \cdot Y_F(i).$$

At a λ-factor close to 1, an insignificant improvement of 0.5% absolute (compared to the recognition of the spelled part only) was observed. However, if the N-best list $Y_F(i)$ for the fluently spoken names is only computed for those names which were already found in the N-best list $Y_S(i)$ of the spelled names, the combination of $Y_F(i)$ and $Y_S(i)$ results in a recognition rate of 97.7% names correct at $\lambda = 0.96$, compared to 96.5% on the spelled names only.

The value for λ was determined on a crossvalidation set. Recognition rates for different values of λ are shown in the upper curve in figure 1. The λ-factor close to 1 indicates that the decision is dominated by the letter recognition, which can only be overwritten if the first-best letter hypothesis has only a small safety margin, i.e. is closely followed by competitors with similar scores.

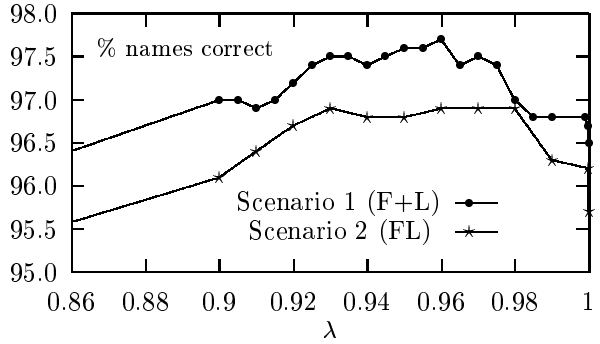

Figure 1: % names correct for a λ-weighted combination of the N-best list of spoken and spelled names (scenario 1 and 2)

3.2. Scenario 2

The task according to scenario 2 is more difficult, because the boundary between the spoken and spelled word is no longer known a-priori. To recognize a spoken and spelled name in one utterance, the dictionary is modified to contain both the pronunciation of the spoken and spelled name in one entry, i.e. "`[Lang_L_A_N_G L A NG - E L - AH - E N - G EH]`" for the name "`Lang`".

Surprisingly, with 86.1% correct, the recognition on the entire utterance is worse than on the spelled part alone! Although the spoken names provides additional acoustic evidence, its less reliable acoustic scores may overwrite a correct decision based on the spelled name alone.

It is possible to adapt a similar approach as in scenario one, by using a λ weighting to strengthen the more robust letter recognition. However, as the length of the spoken and spelled part may differ in each hypothesis, it is no longer meaningful to compare the weighted hypotheses[3]. To circumvent this problem, the boundary of the first best hypothesis was used for the weighting of all hypotheses, resulting in a recognition rate of 89.1%, which is still worse than 93.3% names correct achieved by JANUS on the letter parts only.

To incorporate the MS-TDNN letter recognizer, the spelling segment in the utterance (as identified by JANUS) was re-recognized with the MS-TDNN, resulting in 95.8% names correct. With a more sophisticated approach similar to that described in section 4, 96.9% correct names were achieved after a λ weighting (lower curve in figure 1).

4. LARGE LISTS

Proper names can be recognized like any other words if their pronunciations are added to the dictionary of a speech recognizer. However, if the number of names exceeds the recognizer's maximum vocabulary size (typically somewhere around 65,000 words), a different approach has to be taken.

For very large name lists a two-step approach is employed. First, a coarse recognition run is used to get a reduced list of name candidates. These are then processed in a second pass, in which all the previously described techniques for small word lists can be applied.

The MS-TDNN letter recognizer is able to handle lists of up to 1 million names. Thus, in the case of scenario 1, the list of candidates can be easily reduced if only the spelled names are considered in the first pass.

For scenario 2, we use the JANUS recognizer in a modified version, with only phonemes and letters in its

[3] In that case, the weighted score depends heavily on the length of the spelled part, which is of course undesired.

recognition vocabulary. A special language model (figure 2) enforces that at the beginning of the utterance, only phonemes can be recognized (to account for the fluently spoken name). At some point, the language model switches to the recognition of letters only (to account for the spelled names), hoping for recognitions like "/s/ /m/ /i/ /th/ S M I T H". The corresponding phoneme and letter trigrams were trained using the pronunciations and spelling of all 200,000 last names in the ONOMASTICA dictionary.

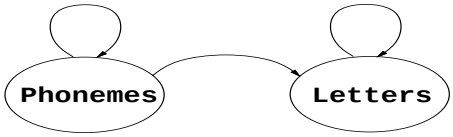

Figure 2: Language modell for the phoneme-letter recognizer

Of course, as opposed to a full dictionary, the recognized sequences of phonemes and letters can not necessarily be interpreted as a legal name, nor is the recognition of the phonemes coupled to the recognition of the letter sequence. However, a list of the most similar 100 or 1000 names can be retrieved from the recognized phoneme-letter sequence. These name candidates are then used in another JANUS recognition run, which is now possible with the full transcriptions, resulting in better boundaries. The letter segments are then re-recognized with the MS-TDNN. Depending on these results new candidates for the fluently spoken names are generated with JANUS. These N-best lists can then be recombined with a λ rescoring as described in the previous section.

We used this technique with the small list of 1337 (for a direct comparison) and with a list of 100,000 unique names. The results of these and the previous experiments are summarized in table 3. Interestingly, the recognition rate for the list of 1337 names improves when the mulit-pass strategy is employed. The reason is that an additional pass is used to re-estimate the boundary between the fluently spoken and spelled part *after* the name list is already reduced in the first pass.

5. FLEXIBLE RECOGNITION

Using the above techniques, the recognizer can be modified so that the user has the choice to spell only, or to speak and spell, resulting in a more flexible system. Both the pronunciation for the spelled only (L) and the spoken and spelled (FL) name are added to the dictionary. An input of either L or FL can be distinguished with almost 99% correct, resulting of 95.5% names correct without a priori knowing whether L of FL was spoken. This compares to 96.5% for spelled only recognition.

6. SUMMARY

Spelled names can be recognized with a much higher accuracy than spoken names. By combining the N-best lists of both the spoken and spelled recognition, the overall performance can be improved. However, due to the dominant role of the spelled letter recognition, the combination must be strongly biased towards the spelled letter recognition, and only a relatively modest improvement can be achieved with the additional information provided by the fluently spoken name.

For name lists too large to fit into the recognizer's dictionary, we have sucessfully applied a two-pass strategy, in which a phoneme-letter recognizer is used to cut down the number of candidates. The results are summarized in table 3.

In addition, the examined methods allow for a more flexible recognition. Alternatively spelled only or spoken and spelled names can be recognized almost with the same accuracy as if it is a priori known that only spelled names were used.

Acknowledgments

This research was partly funded by grant 01IV701U7 from the German Ministry of Science and Technology (BMBF) as a part of the Verbmobil project. The ONOMASTICA Database was provided by courtesy of Deutsche Telekom and TU Berlin. The authors would like to thank Alex Waibel and other members of the Interactive Systems Labs for their support and helpful discussions.

7. REFERENCES

[1] M. Betz and H. Hild. Language Models for a Spelled Letter Recognizer. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 856–859, Detroit, MI, May 1995.

[2] R. Cole, M. Fanty, M. Gopalakrishnan, and R. D. T. Janssen. Speaker-Independent Name Retrieval from Spellings Using a Database of 50,000 Names. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 325–328, Toronto, 1991.

[3] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal. The Karlsruhe-Verbmobil Speech Recognition Engine. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 83–86, Munich, April 1997.

[4] H. Hild and A. Waibel. Speaker-Independent Connected Letter Recognition with a Multi-State Time Delay Neural Network. In *EUROSPEECH'93 (3rd European Conference on Speech Communication and Technology)*, pp. 1481–1484, Berlin, September 1993.

[5] H. Hild and A. Waibel. Recognition of Spelled Names over the Telephone. In *Proceedings Fourth International Conference on Speech and Language Processing*, pp. 346–349, Philadephia, PA, October 1996.

[6] J.-C. Junqua, S. Valente, D. Fohr, and J.-F. Mari. An N-Best Strategy, Dynamic Grammars and Selectively Trained Neural Networks for Real-Time Recognition of Continuously Spelled Names over the Telephone. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 852–855, Detroit, Michigan, May 1995. IEEE.

[7] C. A. Kamm, C. R. Shamieh, and S. Singhal. Speech Recognition Issues for Directory Assistance Applications. *Speech Communication*, 17:303–311, Nov. 1995.

[8] B. Kaspar, G. Fries, K. Schuhmacher, and A. Wirth. FAUST - A Directory Assistance Demonstrator. In *EUROSPEECH'95 (4th European Conference on Speech Communication and Technology)*, pp. 1161–1164, Madrid, September 1995.