ROBUSTNESS IMPROVEMENTS IN CONTINUOUSLY SPELLED NAMES OVER THE TELEPHONE

Michael Galler and Jean-Claude Junqua

Panasonic Technologies Inc. / Speech Technology Laboratory
3888 State Street, Suite 202, Santa Barbara, CA 93105, U.S.A.
email: galler@research.panasonic.com and jcj@research.panasonic.com

ABSTRACT

A speaker-independent speech recognizer for continuously spelled names, implemented for a switchboard call-routing task, is analyzed for sources of error. Results indicate most errors are due to extraneous speech and end-point detection errors. Strategies are proposed for improving the robustness of recognition, including tolerance for speech with pauses, and a letter-spotting strategy to handle extraneous speech. Experimental results on laboratory data indicate that with the letter-spotting method, name retrieval error rate is reduced on noisy signals or signals with extraneous speech 60.1%, while it is increased on clean signals from 4.5% to 5.5%. On data collected during a telephone field trial, name retrieval error is reduced 54.1% in offline tests by introducing the letter-spotting algorithm.

1. INTRODUCTION

This paper discusses robustness improvements to an HMM-based speaker-independent continuous speech recognition system for spelled-name recognition over the telephone. The system, described in [1] (see also [2]), is based on a multipass recognition strategy. The first pass employs a bi-gram language model and a Viterbi beam search to produce the N-best (typically 20 hypothesized) letter sequences. In the second pass, the sequences are compared to a dictionary and a dynamic-programming match is used to select the N-best names in the name directory, based on statistics for letter-confusion. A third pass of recognition is then performed in which the acoustic parameters of the signal are fitted to the HMM sequences representing the N-best names from the directory. In this final pass a full Viterbi search is performed on the reduced set of candidate names. In this way an accurate, efficient search is produced by filtering the search space with different constraints through various stages.

The recognizer was integrated with the telephone switch at Speech Technology Laboratory (STL), and several months of field tests were logged in which callers used the system to direct their calls to company staff members. Callers were asked to confirm the correctness of the recognition output with a yes/no response. Calls were recorded, logged, and later transcribed to determine the source of errors.

Figure 1 presents the collected performance statistics. According to the experimental data, the two most common sources of error were pauses interposed in the letter sequence, causing a premature detection of end-of-speech, and extraneous speech, usually at the beginning, e.g. "Smith [pause] s-m-i-t-h". Other major sources of error include low signal energy in data collected through speaker-phones, and other effects of channel mismatch between training and test conditions.

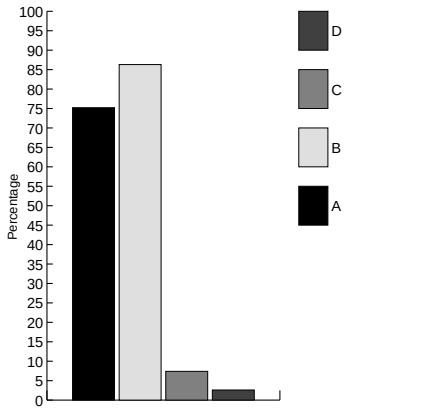

Figure 1. Call Recognition and Error Rates

2. RELATED WORK

2.1. Call Routing

The work described in this paper involves a number of different issues, including call-routing by name recognition, spelled-word recognition, word-spotting techniques, rejection of noise and extraneous speech, and the design of user interfaces for telephone-based speech recognition systems. Speech-operated auto-attendants are now being deployed by many institutions. A number of systems have been introduced experimentally for call-routing that involve speaker-independent, large-directory, whole-name recognition [13], [5], [6]. Some systems have employed hybrid schemes that attempt to manage errors through dialogue, prompts, and user-feedback [7], [8], and a mixture of whole-name and spelled-letter fallback [7]. In contrast with these systems, [9] attempts to perform call-routing for small directories on

small, inexpensive hardware, based on whole name recognition.

The system described here can be seen as a competitor to these systems because it manages name recognition using a small set of acoustic models that run quickly on simple hardware. Also, because of the extra information exploited in the lexical comparison to the dictionary in this multi-pass algorithm, the recognition accuracy is better than for conventional systems which put all the spelled names into a network [1]. It can also be seen as a complementary technology, or fall-back procedure that can be employed by the larger, whole-word systems.

2.2. Robustness Issues

A main issue addressed here is the problem of Out-Of-Vocabulary (OOV) rejection. Although the user instructions are very simple (paraphrase): "Please spell the name of the person you want to call", users have a tendency to pronounce the whole name anyway. Well known techniques exist to deal with this problem, based on the use of filler models and multiple grammars [10], [11]. However, in our work the integration of these techniques with the multi-pass spelled name search is original. Other techniques recently introduced for OOV rejection and keyword-spotting are based on on-line garbage modeling [12]. These techniques have not yet been explored in this work.

3. DATA DESCRIPTION

The database used in our experiments is a subset of the speech telephone corpus collected at Oregon Graduate Institute (OGI) [4]. Over four thousand people called in response to public requests. They were prompted by a recorded voice to spell their first and last names, with and without pauses, together with other information. 60 labeled repetitions of the alphabet (which did not belong to the test set as defined in the OGI CD-ROM) and more than 1200 different calls were selected for the training, 558 calls for the validation and 491 calls for the test. All the calls selected were last names produced without pauses. Moreover, none of the calls contained extraneous speech, line noise or speech related effects such as lip-smack or breath noises (as transcribed in the database).

The three sets selected were subsets of the corresponding training, test and validation sets defined in the OGI CD-ROM. The purpose of the validation experiments was to optimally tune the system's parameters before running it on the test set. As every speaker belongs only to one set (training, validation or test) the experiments conducted are speaker-independent.

491 signals containing correctly spelled names formed the basic test suite. 521 signals containing extraneous speech in which the pronounced name preceded the spelled name were used for a laboratory evaluation of the system. These signals were extracted from the portion of the OGI database where speakers said their last names followed by its spelling. In addition, a set of 159 spelled names from the OGI database containing non-speech utterances and noises was used for laboratory testing.

During the first field trial at Speech Technology Laboratory, 403 calls were recorded over the telephone. Roughly half were in-house transfers using the digital PBX, and half were made from outside the company through the telephone network. These data were also used to evaluate the initial recognition system.

Following further development of the recognition system as described in section 4. and in 5., a new set of field data is being collected to test the robustness-enhanced system. As of this writing, this data set contains 196 recordings of spelled names.

4. ENHANCED ROBUSTNESS BY PAUSE-HANDLING

Many errors were caused by the premature detection of end-of-speech by the Voice Activity Detection (VAD) algorithm. These events were triggered when the speaker hesitated either during the spelling of a name, or between some introductory utterance and the subsequent utterance of spelled letters. The VAD algorithm is implemented as a state machine with 4 states: *non-speech*, *speech-in-progress*, *end-of-speech*, and *false-alarm* (a non-speech event).

State changes are triggered by changes in signal energy, computed adaptively at run-time. Frames of speech data detected by the VAD are passed from the front-end to the recognition module, which computes the forward probabilities frame-synchronously. The backtrack phase of the Viterbi recognition process is triggered by a VAD-detected end-of-speech. In real calls valid speech segments are often interspersed with pauses which, in the original system, caused the VAD to trigger the recognition process to end prematurely.

The spelled-name recognition algorithm was modified to allow signals containing pauses to be recognized. In the modified system, the VAD continues to classify and segment the raw signal as before; however, the recognition module employs a timeout of its own to decide whether an input has terminated. The new algorithm allows up to two seconds of silence between letters before determining the final endpoint was reached. During this interval, it proceeds through the stages of recognition, preparing a tentative output. If the VAD determines the speaker has resumed speaking before the timeout, the second or third pass of recognition aborts and the forward algorithm of the first pass resumes until the next pause is detected. When two seconds of non-speech have elapsed, the tentative response is confirmed and delivered, and recording stops.

In the original field test [3], 7.4% of calls were interrupted due to the problem of slow or interrupted speech. With the system modified as described above, less than 2% of calls in the ongoing trial have been subject to this source of error.

5. MODELING OF EXTRANEOUS SPEECH AND NOISE

5.1. Improved Robustness with Letter-Spotting

A modified recognition strategy was proposed for dealing with the other major source of error. Extraneous speech is managed with a word-spotting strategy (Figure 2.) An initial experiment employed a network with a filler HMM to produce all the N-best sequences of letters for the dictionary alignment. However, for input signals which do not contain extraneous speech this algorithm increased the instances in which the alignment procedure failed to produce the right

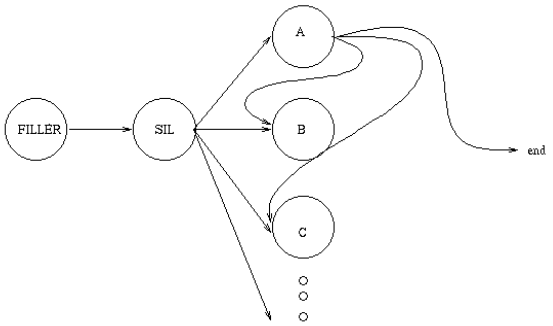

Figure 2. Network for letter-spotting (filler followed by silence, followed by a letter-loop.)

name in its N-best list. This was caused by the increased number of unit errors present in the input sequences, due to the additional complexity of the language model.

A more successful strategy was to fit the signal to two different networks. The first network consists of a filler HMM, followed by silence, followed by the letter models. This network tries to spot the best letter sequence following an initial extraneous noise, word, or phrase. The second network assumes only letters are spoken, and tries to fit the whole signal to the best letter/pause sequence. In both cases the transitions between letter HMMs are weighted by bigram probabilities. Each of these two recognition passes produces an N-best list of sequences. In the next stage of recognition, the sequences are aligned against the name dictionary to retrieve the most likely names with respect to the confusion statistics on letters.

Two alignments are made to compare letter sequences to the names in the dictionary. Since the network without a filler model is used, the input to one alignment procedure is guaranteed to include the best sequences that the letter models can produce for a signal not containing extraneous speech. If the signal does contain extraneous speech, the network containing filler is more likely to produce a sequence acoustically similar to the actual spelled name, and the input to the second alignment procedure will contain that sequence. In our experiments, the output of each alignment produces 10 hypothesized names, and these are combined into a list of 20 candidates. In the final pass a dynamic grammar is generated consisting of the best candidates from the dictionary, with an optional initial filler model. A Viterbi acoustic recognition pass is performed using this grammar to select the most likely name.

5.2. Filler Models and Networks

Different kinds of filler models were constructed and tested. One consisted of a single speech unit trained on all non-silent segments of the NTIMIT 8kHz training corpus, followed by a silence unit. The filler unit was a left-to-right model containing 8 states, with 8 gaussian distributions per mixture density. The grammar containing this sequence segments the input utterance into an initial, extraneous burst of speech, followed by a silence, and ending with a sequence of letters. This network is motivated by analysis of the error data, in which a spelled name was often preceded by an introductory word or phrase, and then by a pause of some duration.

Data Set	Size	Original System Error	Letter-spotting System Error
names only	491	4.5%	5.5%
names with extr. speech	521	75.8%	29.8%
names with noise	159	28.3%	11.3%
STL field trial data	403	23.3%	10.7%

Table 1. Name Retrieval Experiments – Results

The second kind of filler tried was a looped-phoneme network trained on NTIMIT. This network modestly increased the number of correct outputs from dictionary but at a cost of increased computation. Because of the success of the simple filler model as seen below, this approach was dropped.

5.3. Specific Noise Models

A final variation on the filler network was tried, in which the filler models described above were replaced with a small set of specific noise models, lip-smack, breath-noise, and line-noise, trained on the OGI database. These 4-state, 16 distribution models were added to the letter-recognition network, and tested for accuracy in detection of extraneous noise or speech in the spelled-name data. The performance of the filler-based letter-spotting was compared to this method of using models trained for specific noises.

6. EXPERIMENTAL RESULTS

The results for the original and modified name recognition algorithms are summarized in Table 1. The results are presented as name-retrieval error, or the percentage of recordings for which the wrong name was selected at the end of the final pass.

The original system was shown to achieve a 4.5% retrieval error rate (on 491 OGI test signals) for a 3,388 word dictionary. When tested on a separate subset of 521 OGI signals consisting of extraneous-speech *plus* spelled-name data, the same system had an error rate of 75.8%. When the letter-spotting algorithm was tested on the spelled names, recognition error increased from 4.5% to 5.5%. However, on the extraneous speech data, the error rate decreased from 75.8% to 29.8%. On the noisy signals, there was a 60.1% reduction in error rate.

Of particular interest was the comparison between the filler model and the noise model performance on data containing noise, as seen in Table 2. As should be expected, the filler performed much better than the noise models at detecting extraneous speech. But the filler model performed equally well in name retrieval as the noise-modeling network in detecting speech with noise. However, better performance may be achievable with more precise noise models.

7. CONCLUSIONS

Through the careful examination of the performance of a speech recognition service on live data, it is possible to iden-

Data Set	Letter-spotting with Filler Model	Letter-spotting with Noise Models
names with extr. speech	29.8%	56.0%
names with noise	11.3%	11.3%

Table 2. Error rates of letter-spotting by filler and by noise models

tify the largest sources of error. Often these problems can be greatly alleviated by small adjustments to the user interface, and the simple application of existing techniques.

The Speech Technology Laboratory call routing algorithm has been improved by the application of word-spotting and noise modeling techniques in a way that significantly increased the robustness of the live system. Using a noise-independent filler model the system outperformed a system using specific noise models, especially when all kinds of extraneous inputs were taken into account. The filler model approach has the considerable advantage that specific noise models do not have to be trained for new environments and recording conditions.

Future work will include modifying the system to perform rejection, and to ignore extraneous speech at any point during an input utterance. The system will be enhanced to handle multiple calls simultaneously in real-time and field-tested on a much larger user base.

REFERENCES

[1] Junqua J.C. et al., "An N-best strategy, Dynamic Grammars and Selectively Trained Neural Networks for Real-Time Recognition of Continuously Spelled Names Over the Telephone", *ICASSP*, May 1995.

[2] Junqua J.C., "SmarTspeL: A multi-pass recognition system for name retrieval over the telephone", *IEEE transation on speech and audio processing*, (to appear in forthcoming issue).

[3] Junqua J.C., and Galler M., "Performance Evaluation of SmartTspelL: A Continuou sly Spelled Name Recognizer over the Telephone", *IVTTA*, September 1996.

[4] Cole R., Roginski K., and Fanty M., "English Alphabet Recognition with Telephone Speech", *Eurospeech 91*, pp. 479–482.

[5] Yamamoto et al., "A Voice-activated Telephone Exchange System and its Field Trial", *IVTTA*, 1994.

[6] Billi R. et al., "Interactive Voice Technology at Work: The CSELT Experience", *Speech Communication*, 17, pp. 263–271, November 1995.

[7] Johnston D., Whittaker S.J., and Attwater D.J., "An Overview of Speech Technology for Telecom Services in the United Kingdom", *IVTTA*, pp. 12–15, September 1996.

[8] Kellner A., Rueber B., and Seide F., "A Voice-Controlled Automatic Telephone Switchboard and Directory Information System", *IVTTA*, pp. 117–120, September 1996.

[9] Fraser N.M., Salmon B., and Thomas T., "Call Routing by Name Recognition: Field Trial Results for the Operetta System", *IVTTA*, pp. 101–104, September 1996.

[10] Rose R.C., Juang B.H., and Lee C.H., "A Training Procedure for Verifying String Hypotheses in Continuous Speech Recognition", *ICASSP*, pp. 281–284, May 1995.

[11] Rahim M.G., Lee C.H., and Juang B.H., "Robust Utterance Verification for Connected Digit Recognition", *ICASSP*, pp. 285–288, May 1995.

[12] Bourlard H., Dh́oore B., and Boite J.-M., ""Optimizing Recognition and Rejection Performance in Wordspotting Systems", *ICASSP*, pp. 373–376, 1994.

[13] Smith G.W., and Bates M., "Voice Activated Automated Telephone Call Routing", *Proceedings of the Ninth IEEE Conference on Artificial Intelligence for Applications*, March 1993.