ROBUSTNESS ISSUES AND SOLUTIONS IN SPEECH RECOGNITION BASED TELEPHONY SERVICES

Vijay Raman and Vidhya Ramanujam

NYNEX Science and Technology, Inc.
500 Westchester Avenue
White Plains, NY 10604 USA

ABSTRACT

HMM-based algorithms for speaker-dependent recognition as well as speaker-independent recognition form the basis of speech services developed at NYNEX S&T and deployed widely by NYNEX and other telephone service providers. Based on the analysis of the initially deployed VoiceDialing service, robustness of these algorithms was recognized to be a dominant issue. In this paper, we discuss the features of a high-performance, robust speaker-dependent recognition algorithm, and include some deployment issues that were successfully resolved.

I. INTRODUCTION

The initial VoiceDialing service deployed by NYNEX [1] implemented speaker-dependent recognition using special-purpose DTW (dynamic time warping) hardware. Large volumes of customer usage data, including recorded speech data, direct customer feedback and marketing focus group feedback, have been acquired from that first deployment. This has provided us with insight into the robustness issues prevalent in the deployment of telephone-based speech recognition services to a mass market. Algorithmic issues such as out-of-vocabulary rejection, word-spotting, usability from multiple telephones, robustness against background noise, ease of training names, etc., make a significant difference in the usability, and hence acceptability, of the service to the end-user.

A second-generation speech sub-system has since been deployed in which speech recognition algorithms are implemented in software on a general-purpose DSP (digital signal processor). ***HMM*** (Hidden Markov Model) algorithms combined with word-pair grammars have been designed and implemented for this purpose. In particular, the speaker-dependent training and recognition algorithms have utilized the HMM paradigm, and the flexibility of software-based recognition, to provide high-performance speech recognition capability that successfully addresses many of the shortcomings identified in the prior VoiceDialing technology. The algorithms/models developed to this end are discussed in Sections II, III and IV.

The VoiceDialing service has also been grown to incorporate speaker-independent recognition (network control commands, among other applications) [2] on the same tier as speaker-dependent recognition, and this requires further enhancements in rejection capabilities. This is described in Section III.

A number of cellular-telephone users currently access the VoiceDialing service from either their home or cellular telephones, and this number is anticipated to grow significantly. Robustness against noise has been addressed through the design and implementation of a fast noise-canceller, as in Section V.

The recognizer has been designed to be highly parameterized to allow site or user-specific modifications. This has proved useful in optimizing for performance. Recognizer performance, tuning and trade-offs are discussed in Section VI.

A deployment issue which arose in our deployment indicates a problem of general concern to service providers. Specifically, the deployment of the new recognition algorithms required a transparent transformation of the "old" user's models into "new" models. More generally, as deployment coverage of services such as VoiceDialing and other speaker-dependent/speaker-adaptive/peaker-verification increases, there will be requirements for technology changes, vendor changes, and situations such as VoiceDialing-directory *portability* there will be the need for methods to exchange user-data (models) between differing recognition systems, and the recognizer capability itself must be expanded. Algorithms developed at NYNEX to allow transparent change-over to the new system (which has different front-end signal processing, and HMM modeling) are discussed in Section VII.

II. IN-VOCABULARY MODELING

The front-end generates 10 LPC-based cepstral coefficients, 10 delta-cepstra, energy and delta-energy, at the frame rate of 20 ms, and a window size of 30 ms. All features are quantized to 8 bits. The delta-cepstra were found

to have minimal effect on in-vocabulary discrimination, but affected out-of-vocabulary rejection favourably.The specific form of delta-cepstra calculation is with a standard interpolation formula over 5 frames, for compatibility with speaker-independent recognition; however, shorter windows did not adversely affect performance.

The recognition network is constructed from the specified grammar and continuous Gaussian-mixture HMM models. Viterbi decoding is carried out with a beam search.

The ***HMM model*** for a spoken name is constructed from two utterances, although any number may be used. The number of states is quantized to one of three values, based on the duration of the utterances; these values were chosen as a compromise between the requirements of mean estimation and temporal resolution. A seed model is built from the first utterance, and Viterbi alignment of the second utterance is used to update the model. A single-mixture Gaussian leads to the Euclidean distance metric.The use of a pooled variance from the speaker utterances was found to reduce performance when the data was small, and hence global variance normalization was used.

During training, the algorithm determines whether (a) the two repetitions are acoustically consistent, and (b) the name is not acoustically similar to an existing name in the directory. For this, a recognition pass is run during the training process. Likelihood scores are used to determine if the above conditions are met.

The ease of training names (without being rejected as similar to an earlier one) has been considerably enhanced with respect to the prior technology, without sacrificing discrimination. The average number of names successfully trained from a field database increased by about 25% compared to the prior system. Live testing upto the maximum of 75 names confirms this improvement.

III. OUT-OF-VOCABULARY REJECTION AND WORDSPOTTING

A. Models and Grammar

The VoiceDialing user may utter names which are not in his directory, not remembering whether they were trained or not. This can occur with less frequently used names. In the situation where a family shares a single VoiceDialing directory, multiple users can aggravate a false-acceptance problem

A word-pair grammar is built on-line during recognition or training, utilizing speaker-specific information. The models for competing names trained by the user are on parallel arcs in this grammar. Additionally ***three speaker-dependent garbage models*** and ***one speaker-independent garbage model*** are placed on arcs parallel to the names, to provide out-of-vocabulary rejection.

The three speaker-dependent garbage models have, respectively, number of states equal to one of the three quantization bins to which utterances are assigned during training. Each of these garbage models is computed on-line as a composite of the names in its quantization bin. Noting that there may not be any names in that bin, each of the garbage models is seeded from the first utterance spoken by the user, using an averaging scheme, and is updated subsequently as needed. Consequently, these garbage models act as duration-wise out-of-vocabulary models.

The probability attached to the grammar arcs for the speaker-dependent garbage models is linearly adjusted according to the number of names contributing to the model, to avoid over-representation of a name in a garbage model. Figure 1 gives histograms of scores for utterances recognized as in-vocabulary, in a database test. Garbage models are effective in reducing false accepts without a large penalty in correct accepts. Furthermore, in all cases, a score threshold (in this case around 10.0) is effective in reducing false accepts further by about 50% for a very small in-vocabulary penalty.

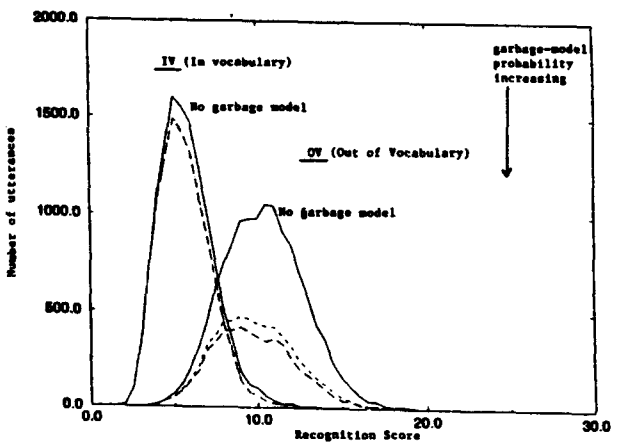

Fig. 1. Histograms of scores for in-vocabulary and out-of-vocabulary utterances recognized as in-vocabulary, as a function of the probability assigned to the speaker-dependent garbage models.

The single *speaker-independent garbage model* was built from a large number of isolated utterances collected over the telephone network. This models has further enhanced out-of-vocabulary rejection, additional to the other rejection strategies used.

The situation frequently arises wherein the user has trained the first name (e.g. "John") but forgets and says the full name ("John Smith"). Additionally, other peripheral speech, household conversation, radios are observed in VoiceDialing usage, and calls for ***word-spotting***.

A word-spotting garbage model is built as a single-state composite of the multiple speaker-dependent garbage model, effectively forming a speech model for the speaker. Grammar arcs with this model precede and succeed the tar-

get-speech arcs. Thus target utterances spoken with preceding or succeeding speech are recognized correctly. Additionally, this has proved effective in capturing other peripheral sounds.

B. Application-Specific Models and Grammar

Currently, the VoiceDialing service is being offered wherein the speaker may say a name in his directory or a network keyword such as *Call Forwarding, Repeat Dialing*, etc. This is accomplished by parallel speaker-dependent and speaker-independent recognition. Given the occurrence of such specific known phrases, the robustness of recognition is enhanced as follows.

Application-specific garbage models were built from collected telephone speech, with the model parameters set as for speaker-dependent models. The phrases modeled were complete phrases, such as *Call Forwarding*, or components, such as *Voice* (part of *Voice Dialing* and *Voice Mail*). 6 models were built due to memory restrictions. The speaker-dependent grammar places these in parallel to the target utterances.

Note that because of the small model-parameter size, detection of keywords does not achieve high accuracy but the rejection of application-specific keywords was improved significantly. The probability attached to these garbage models is configurable as a system parameter.

A feature of the current VoiceDialing service is that a user can say *"**Call Forwarding to <name>**"*, and have this recognized (***<name>*** being in his directory) and appropriate action taken. The word-spotting grammar was modified to include the model for an application-specific phrase (in this case, *Call Forwarding*) on an arc preceding the target utterance.

IV. BACKROUND MODELING

Background, or silence, modeling is important to improving alignment of incoming speech and models. To this end, ***3 single-state silence models*** were built from a body of collected data. The training data was coarsely classified by average energy level, and the data used to build silence models for these environments.

It was observed, however, that Viterbi alignment could still be significantly inaccurate for a specific switch, telephone-set, etc. ***Dynamic silence modeling*** was then introduced Since speech detection is used to initiate recognition, the signal data prior to speech onset represents the actual ambience; this data is used to build a silence model on-line, whenever it is sufficiently long in duration. This model then operates in parallel with the fixed silence models. In fact, in actual usage, there is indeed an adequate period before speech onset, and the dynamic silence model is created. In laboratory tests, the dynamic silence model was found to most often be part of the Viterbi best path, and was therefore judged effective.

V. NOISE CANCELLATION

An automatic noise-classification and cancellation algorithm for stationary noise was described originally in [3]. This approach has formed the basis of a noise-canceller which has currently been incorporated into the front-end. This capability is to be deployed shortly, and testing results given in other sections do not include noise-cancellation.

Testing conducted on land-line data and mobile data has demonstrated a reduction of around 30% in error rate under noisy conditions with a negligible change in performance under quiet conditions.

VI. RECOGNIZER EVALUATION

Recognition and training is driven by a set of configurable parameters that can be modified according to the needs of an application, or to address site-specific or customer-specific issues.

Parameters that are used to modify the grammar and the Viterbi search: grammar probabilities to be attached to the speaker-dependent garbage models, the application-specific garbage models, and the word-spotting garbage models, pruning beam width, etc.

Parameters that are used to control post-processing: recognition score threshold for rejection, various training score thresholds for testing the similarity to existing names, and the consistency of the two training utterances, duration limits for very short or long utterances, and parameters for testing Viterbi alignment.

Performance of the system was measured/analyzed using internal databases, field trials and live user testing in the laboratory.

The dominant database used, denoted the ***CVDIAL*** database, consists of 77 speakers, recorded over a T1 data-collector at NYNEX Science and Technology. Each of them completed upto 10 sessions, in which they read from a list of 15 names, and 15 network command phrases, such as *Voice Messages*. The names constituted in-vocabulary items, and the network commands out-of-vocabulary items. Training used two different repetitions of each name, repeated in round-robin fashion over the full set. Recognition performance was averaged over all tests. Tests performed on other databases are not described here.

In-vocabulary and out-of-vocabulary performance was obtained for various parameter configurations representing different levels of rejection. The probabilities attached to the speaker-dependent garbage models and those attached to the application-specific garbage models were varied to obtain the performance curves shown in Figure 2.

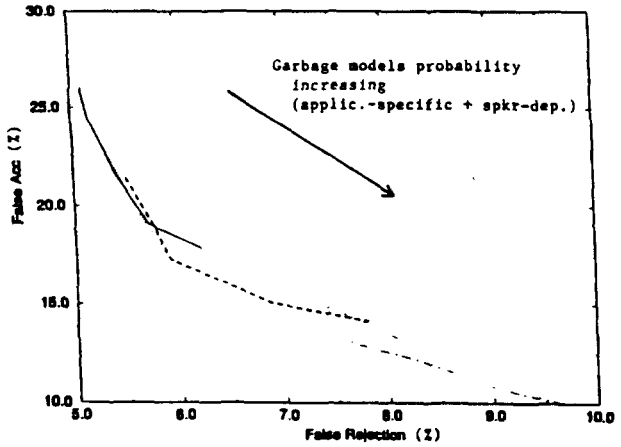

Fig. 2. Performance as a function of the probability of speaker-dependent and application-specific garbage models. Each sub-curve is for a fixed probability of the application-specific models.

Analysis of the user data from a field trial conducted in the NYNEX region indicated that the performance of the recognizer in the deployed system was consistent with the results predicted from the database testing described above. Customer feedback from the field trial further indicated that there were improvements in usability (including from multiple telephones in the household).

A point of some concern is that although overall robustness of recognition was improved significantly, the distribution of performance across speakers, Figure 3, indicates that even with an average in-vocabulary performance in the 90's, about 8% of the population has performance below 80%, reducing usability of the service.

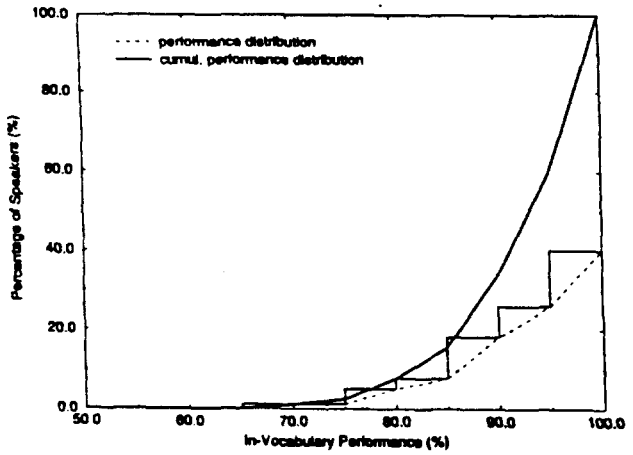

Fig. 3. In-vocabulary performance distribution over speakers.

VII. TEMPLATE CONVERSION

NYNEX and other VoiceDialing service providers using the NYNEX platform had a large customer base of subscribers who had trained their directories using the DTW technology deployed initially. A marketing-critical issue was that of upgrading to the new platform and HMM recognition technology in a transparent fashion (viz. not requiring the customer to re-train the names in his directory). Algorithms were therefore developed to this end.

It was recognized that there were effectively three pieces of speech information on each name trained by the user: two DTW templates, and one confirmation recording. The recording was in 16K ADPCM format. The DTW templates used cepstral features quantized to 4 bits each, at a frame rate of 22.5 ms. The methodology employed devised suitable dequantization tables, and generated HMM models from the DTW templates and the confirmation recording. Garbage models were built from the set of name models.

Use of the confirmation recording in training was found to improve performance significantly. In database tests, shown in Table I, the recognition error rate decreased by 50% or more due to the confirmation recording. Dequantization tables were also optimized for performance.

The coarseness of the feature quantization, and the differing frame rates of the DTW and HMM systems leads to a poorer match between the incoming speech and the converted models then with HMM-trained models. Compensation for this has to be done during the Viterbi search, and was accomplished by a frame-wise score modification for the models of converted names.

The conversion process has been run successfully at a number of sites, covering a large number of customers.

VIII. SUMMARY

HMM modeling in conjunction with word-transition grammars has provided the framework for the development of speaker-dependent and speaker-independent garbage models and background/silence models. The improved rejection and word-spotting capability is essential to the robustness of speech recognition in typical user environments. A deployment issue regarding introducing the improved algorithms raises a problem likely to be of wider interest as service deployment increases, namely the exchange of user-specific recognition models

REFERENCES

[1] G. Vysotsky, "VoiceDialing - the first speech recognition based telephone service delivered to the customer's home," Proc. IVTTA, Kyoto, 1994.

[2] A. Asadi, D. Lubensky, L. Madhavrao, J. Naik, V. Raman, G. Vysotsky, "Combining Speech Algorithms into a "Natural" Application of Speech Technology for Telephone Network Services,", Proc. Eurospeech 95, Madrid, September 1995.

[3] V. Raman and J. Naik, "Noise Reduction for Speech Recognition and Speaker Verification in Mobile Telephony," Proc. ICSLP, Yokohama, 1994.