SPEAKER-INDEPENDENT NAME DIALING WITH OUT-OF-VOCABULARY REJECTION

C.S. Ramalingam *Lorin Netsch* *Yu-Hung Kao*

Texas Instruments, Inc., Personal Systems Laboratory
P.O. Box 655303, MS 8374, Dallas, TX 75265, USA

ABSTRACT

In this paper we propose a system for speaker-independent name dialing in which a name enrolled by a user can be used by other members in a family or co-workers in an office. We use speaker-independent sub-word models during enrollment; the recognized sub-word string is later used during recognition. We also present a mechanism for rejecting out-of-vocabulary (OOV) phrases. The best in-vocabulary (IV) correct and OOV rejection performance for other speakers is 90%/60% (IV/OOV) on a database containing eighteen speakers. If the orthography is known, the best performance is 96%/65%.

1. INTRODUCTION

There is a growing demand in the telecommunication marketplace for speech-recognition based user customization. For example, the user wants to be able to say "Call John Smith", whereupon the system is expected to recognize the name and dial the appropriate number. To accomplish this, the user has to first enroll a set of names, at which point the system creates a model for each name in the set. The enrolled models are used later for recognition. The models could be either speaker-dependent or speaker-independent. In general, speaker-dependent models are more accurate; but, as the number of enrolled names increases, the memory required for storing the models becomes prohibitively large. To overcome this problem, we use speaker-independent, sub-word models for name recognition. This means that one need not store acoustic models for each name, resulting in significant savings in memory. This approach has the added attraction that it can handle another growing demand of residential customers, viz., the ability of one family member to use a name enrolled by another.

The idea of using speaker-independent models to generate speaker-dependent templates was first developed and tested by Scruggs, Wheatley, and Ittycheriah at Texas Instruments in 1991 (unpublished report). A similar idea was outlined in an extended abstract by Buhrke, et al. at the 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications [1]. Jain, et al. [2] rediscovered this approach and used it for voice dialing. Thus far, this idea has been viewed as creating *speaker-specific* templates using speaker-independent models. In this paper we point out that the templates can be used by other speakers as well, with acceptable level of performance. We also describe a method for achieving out-of-vocabulary (OOV) rejection and present results when the system is used by (a) same speaker, and (b) other speakers.

2. ENROLLMENT AND RECOGNITION ALGORITHMS

2.1. Enrollment

The enrollment string for a name spoken by a user is obtained by recognizing it using a phonotactic grammar. The recognizer is based on Hidden Markov Models (HMMs) [3]. The phonotactic grammar specifies the allowed phone sequences in terms of the possible sound sequences in the language. The basic sound units used in our grammar are forty-six monophones. To increase accuracy, each monophone is mapped into all its context-dependent phones. The acoustic models used in all our experiments are sex-segregated. The recognizer outputs a sequence of labels identifying the phone models and the non-speech models (initial and final non-speech models are excluded) for each enrollment. This sequence is used later as a "grammar" for recognition. As an example, for the word "Mom" spoken by a male speaker, the recognizer could output the sequence "`m_m aa_m m_m`", which is later used for recognition.[1] The suffix "`_m`" signifies these are male models. Because the phone models are speaker-independent, the above enrolled sequence can be used by other male speakers as well. The accuracy of the recognition will, of course, depend upon the closeness of the pronunciation. To accommodate female speakers, we

[1] For simplicity, the illustrative examples are given in terms of monophones.

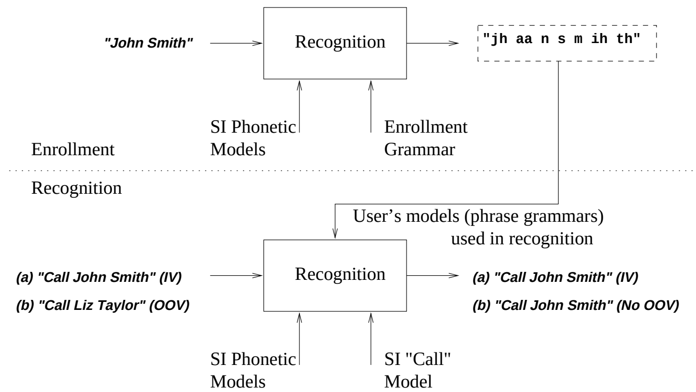

Figure 1. Enrollment–recognition scheme using speaker-independent phonetic models for voice-dialing. In the above example, only one name has been enrolled to illustrate the lack of OOV rejection. (SI = speaker-independent, IV = in-vocabulary, OOV = out-of-vocabulary.)

do not demand an additional enrollment (by a female speaker), but derive one from the existing string by replacing the suffix "_m" with "_f" to get "m_f aa_f m_f"; this is used as an additional grammar during recognition. If the acoustic models are not sex-segregated, deriving the opposite-sex enrollment string will not be necessary.

In an actual application, the user may use the optional anchor word "Call", in which case the grammar should be modified to handle this.

2.2. Recognition

The speaker-independent phone models along with the grammars obtained during enrollment are used for recognition. An enrolled name can be spoken by either the same speaker or by someone else, because the sub-word units are speaker-independent. Implicit in this statement is the assumption that the pronunciations are reasonably close. The complete enrollment–recognition scheme is shown in Figure 1.

In a practical system, a user is quite likely to utter phrases that are not on his list. The system shown in Figure 1 is incapable of rejecting out-of-vocabulary (OOV) names. Misrecognition rather than rejection can be a very costly error in a telephone-dialing system because it will result in dialing the wrong number. We now describe a method for rejecting OOV phrases.

3. OUT-OF-VOCABULARY (OOV) REJECTION

The enrollment grammar with a penalty is added in parallel to the grammars associated with the enrolled phrases. Observe that the grammar associated with an enrolled phrase is nothing but a specific path along the enrollment grammar. When an IV phrase is uttered during recognition, two virtually identical paths are available to the recognizer: one along the grammar that was obtained during enrollment for this phrase, and the other along the enrollment grammar itself (since it is now available in parallel). However, because the latter has a penalty attached to it, it will be discarded during Viterbi decoding. Now consider an OOV phrase. None of the enrollment grammars will match well (the degree of mismatch will depend on how different the OOV phrase is from the IV phrases). On the other hand, the parallel enrollment grammar will be able to parse this phrase and match it better despite the penalty that goes with it. Hence this input is declared as an OOV phrase (see Figure 2). The above approach is reminiscent of the method used by Asadi, et al. [4].

The value of the penalty can be adjusted to yield a desired amount of OOV rejection percentage. The ability to reject OOV phrases will always be at the expense of IV recognition performance. However, if the drop in IV accuracy is only slight, but significant OOV rejection is still possible, the above approach would be viable (as borne out by the experiments described later).

4. TEST CORPUS: HANDSET DATABASE 2

The speech corpus used in the experiments described in this paper is the Handset Database 2 (HSDB_2) collected at Texas Instruments. Speech was collected from eighteen speakers (eight male, ten female), each of whom uttered thirty names. Ten of these names were common to all speakers; the remaining twenty names were unique to each speaker. The list consisted of both short names (e.g. Mom) and long ones (e.g. Jane Duderstadt). Each name was repeated thrice for enrollment. In our experiments we used only two tokens. For testing, the names were collected with and without the anchor word "Call" preceding each name. The handsets used were carbon button, electret, cordless, speaker-phone, cellular hand-held, and hands-free cellular. To reduce the number of experiments, the combinations of training and test-

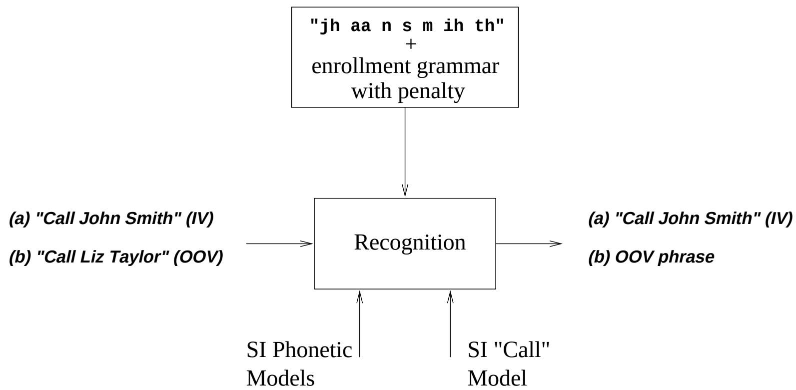

Figure 2. Modified recognition system that is now capable of OOV rejection. The enrollment grammar is added in parallel with a penalty; ideally, this path is taken for OOV phrases.

ing conditions were restricted to electret, cordless, and cellular hand-held. We chose these three handsets because they are widely used.

5. EXPERIMENTS AND RESULTS

For each speaker, two tokens were used for enrolling each of the thirty names and hybrid grammars ("call" optional) formed from them. In initial testing, we found that the phonotactic grammar was consuming a lot of resources (CPU time and memory) during enrollment. Therefore, we replaced it with a bigram grammar [5], which considerably reduced enrollment resource usage. The test data consisted of both "call" and "no call" sentences. The combination of training and testing conditions were drawn from three handsets: electret (ELE), cordless (COR), and cellular hand-held (CHH). The following experiments were performed for each speaker.

Same-speaker IV performance: Each speaker's thirty-name grammar was loaded and tested with that speaker's test data. The average number of test sentences for the electret handset was 180; for cordless and cellular hand-held, it was 120.

Same-speaker OOV performance: Grammars corresponding to the ten common names were loaded and tested with the sentences containing the remaining twenty names from that speaker. The average number of test sentences was 120.

Other-speakers IV performance: All thirty names were active; the test data were the ten common names spoken by all the remaining speakers. The average number of test sentences for the electret handset was 1020; for cordless and cellular hand-held, it was 680.

Other-speakers OOV performance: All thirty names active; the test data were OOV names uttered by all the remaining speakers. The average number of test sentences was 1900.

The results are summarized in Table 1. The numbers represent the percentage of in-vocabulary correct and out-of-vocabulary rejection, and are the average over all eighteen speakers. In these calculations, call-deletion errors have been ignored. That is, if the optional "call" spoken preceding a name is deleted by the recognizer but the name is correctly recognized, the result is not treated as an error.

For comparison purposes, known orthography experiments were also carried out. For each name in a speaker's list, its enrollment string was constructed by an expert phonetician rather than deriving it by recognizing the corresponding enrollment phrase. For example, the enrollment string associated with "`Hank Hayes`" was "`hh_m ae_m ng_m k_m hh_m ey_m z_m`" (along with its female counterpart). Note that in the known orthography experiments, there is no distinction between "same speaker" and "other speakers". However, because ten names are common, their orthographies were tested by all speakers; whereas, for the non-common names the number of test sentences were limited to the speakers that said those names. Both in-vocabulary and out-of-vocabulary performance were studied. The results for the three handsets are given in Table 2.

6. CONCLUSIONS

Table 1 shows that good "other speakers" performance is possible because of the speaker-independent nature of the sub-word models. The mechanism described in Section 3. provides reasonably good out-of-vocabulary rejection, but still needs improvement. The same-speaker OOV rejection is higher than other-speakers OOV because it was tested with only ten names active (as opposed to thirty), due to database limitations. If thirty names were active, this number would become lower. Although the best performance is obtained when the training and testing handsets are electret, the performance of the cordless handset is quite close. The cellular hand-held results are significantly poorer, and need additional signal processing to improve performance. Currently, we

Train	Test	Same–IV	Same–OOV	Other–IV		Other–OOV	
				SS	OS	SS	OS
ELE	ELE	98.3	73.6	90.4	88.0	61.5	60.9
	COR	96.3	72.4	88.6	86.9	57.7	57.5
	CHH	87.5	73.9	77.9	74.2	59.9	59.4
COR	ELE	95.8	74.0	86.4	84.0	62.8	62.9
	COR	97.9	62.2	86.7	83.9	56.9	57.7
	CHH	87.3	69.8	78.7	74.5	58.3	58.3
CHH	ELE	90.7	77.1	79.1	77.9	65.5	64.8
	COR	92.0	71.0	81.7	79.5	57.2	58.3
	CHH	90.0	59.3	77.5	73.5	53.5	54.1

Same = same speaker
IV = in-vocabulary
SS = same sex
Other = other speakers
OOV = out-of-vocabulary
OS = opposite sex

Table 1. SIND performance summary for the HSDB_2 corpus (electret, cordless, and cellular hand-held).

are working on increasing other-speakers IV accuracy and also other-speakers OOV rejection (which place conflicting demands on the system) across all handsets.

Acknowledgements

The authors wish to thank Jack Godfrey of the Speech Group for his help with the "known orthography" experiments.

Test	IV	OOV
ELE	96.2	64.8
COR	95.4	59.4
CHH	88.0	58.8

Table 2. "Known Orthography" results for the handsets electret, cordless, and cellular hand-held.

REFERENCES

[1] E. R. Buhrke, T. Jacobs, and J. Wilpon, "A comparison of algorithms for speaker-dependent speech recognition for network applications," in *First IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, (Piscataway, NJ), October 1992.

[2] N. Jain, R. Cole, and E. Barnard, "Creating speaker-specific phonetic templates with a speaker-independent phonetic recognizer: Implications for voice dialing," in *Proceedings of IEEE ICASSP–96*, vol. I, (Atlanta, GA), pp. 881–884, Apr. 1996.

[3] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ: Prentice-Hall, 1993.

[4] A. Asadi, R. Schwartz, and J. Makhoul, "Automatic detection of new words in a large-vocabulary continuous speech recognition system," in *Proceedings of the IEEE ICASSP–90*, (Albuquerque, NM), pp. 125–128, April 1990.

[5] J.-C. Junqua and J.-P. Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications.* Boston, MA: Kluwer Academic Publishers, 1996.