STUDIES IN MASSIVELY SPEAKER-SPECIFIC SPEECH RECOGNITION

Yu Shi and Eric Chang

Microsoft Research Asia

{yushi, echang}@microsoft.com

ABSTRACT

Over the past several years, the primary focus for the speech-recognition research community has been speaker-independent speech recognition, with the emphasis of working on databases with larger and larger number of speakers. For example, the most recent EARS program which is sponsored by DARPA calls for recordings of thousands of speakers. In this paper, however, we are interested in making speech interface work well for one particular individual. For this purpose, we propose using massive amounts of speaker-specific training data recorded in one's daily life. We call this *Massively Speaker-Specific Recognition* (MSSR). As a pre-research, we leverage the large corpus we have available from speech-synthesis work to study the benefit of MSSR only from acoustic-modeling aspect. Initial results show that by changing the focus to MSSR, word error rates can drop very significantly. In comparison with speaker-adaptive speech recognition system, MSSR also performs better since model parameters can be tuned to be suitable to one particular individual.

1. INTRODUCTION

Over the past few years, the primary focus for the speech- recognition research community has been *speaker-independent speech recognition* (SISR), with the emphasis of working on databases with larger and larger number of speakers. For example, the most recent EARS (*Effective, Affordable, Reusable Speech-to-text*) program sponsored by DARPA calls for recordings of thousands of speakers. This type of systems is desirable to applications where the amount of speaker-specific training data is not sufficient, say less than several hours. The traditional method to obtain an ideal *speaker-dependent speech-recognition* (SDSR) system is to adapt an SISR system to the specific speaker by using a small training set. In this case, a *speaker-adaptive speech-recognition* (SASR) system may outperform both SISR and SDSR systems. The latest work comparing the SISR, SDSR, and SASR systems we could find is the paper done by Lee and Huang [1], where the SDSR system obtained approximately 50% of the *word error rate* (WER) of an SISR system and the WERs with SASR systems are always equal to or lower than that with SDSR systems. In their paper, however, the SDSR systems were trained on waveforms of at most 2 hours long. But how about increasing the training data to massive amounts? With the increasing popularity of household and professional voice/video-recording equipments and voice-recording capability built into personal electronic devices such as mobile phones, it is relatively convenient to get massive amounts of speaker-specific speech in daily conversations. Therefore, we are interested in using these data to train SDSR models so that the performance of speech interface is optimized for one particular individual. We call this *Massively Speaker-Specific Recognition* (MSSR). In this study, as a pre-research, we leverage the large corpus we have available from speech-synthesis work to investigate the benefit of using massive amounts of speaker-specific training data. Initial results show that by changing the focus from SISR to MSSR, WERs can drop very significantly. We also compare the effect of performing MSSR with SASR in this study.

The rest of this paper is organized as follows. Section 2 gives a brief comparison between MSSR proposed here and the recommend speaker-enrollment process in most commercial dictation software systems. Experiment platform and speech corpus used by the authors are described in Section 3. This is followed by a more detailed description of training-subset-selection process in Section 4. Section 5 is devoted to demonstrating the effectiveness of MSSR. The paper is summarized in Section 6 with future research directions being proposed.

2. COMPARISON WITH SPEAKER ENROLLMENT

The traditional approach to tailoring an SISR system to a particular speaker have included a wide variety of speaker-adaptation techniques, such as MLLR (*Maximum Likelihood Linear Regression*) adaptation, MAP (*Maximum A Posteriori*) adaptation, *speaker-adaptive training* (SAT) [2], and speaker-selective training [3] [4].

While most of commercial dictation software systems such as NaturallySpeaking, ViaVoice, and speech-recognition capability built into Microsoft OfficeXP and TabletPC all recommend speaker enrollment, the methodology used is typically reading prepared scripts attached to the program. We have found that while users can be taught to speak somewhat clearly and distinctly during the enrollment process, in actual usage, it is difficult to maintain such careful speaking style.

Another challenge in speech recognition is the mismatch between the lexicon and the language model of the system and the user's speech. In recent years, however, many promising work have been carried out in this area. For example, many commercial dictation systems allow the user to import such common written text so that both the lexicon and the language model can be updated. Similarly, there has been a great deal of work of leveraging the text available on the Internet to further boost the coverage of the lexicon and the language model [5], [6]. Like the acoustic-model enrollment, texts inputted into the systems are usually much

0-7803-8484-9/04/$20.00 ©2004 IEEE

more formal than colloquial ones.

In MSSR, we are proposing that each speaker's spontaneous speech is recorded in his/her daily life. The recorded speech can then be automatically transcribed and manually verified. Based on the transcribed speech, one is easily able to train speaker-specific acoustic model where parameters are well tuned, while with the text-format transcriptions, the lexicon and language model which match the speaker's habit can be easily created as well. Since the speech of the particular speaker are captured in normal daily conversations, the speaker's speaking style can be truly reflected. While automatic transcription may not be very accurate, it is relatively affordable to have some speech transcribed [7].

3. EXPERIMENT CONDITION

In this paper, the experiment platform is set to the reference English speech-recognition engine v7.0 in Microsoft *Speech API* (SAPI) 6.0 SDK. It is a high-performance speech-recognition engine aimed at both server and PC deployment and capable of *Large-Vocabulary Continues Speech Recognition* (LVCSR) and *context free grammar* (CFG) based recognition.

In this paper, we investigate the performance of LVCSR of all SISR, SASR, and MSSR systems. Whole-word models are removed from the original engine since we only have regular-word transcriptions available for speaker-specific waveforms. SISR directly experiments on the modified engine. Here we only study the benefit of the acoustic model in MSSR, so all systems use the same trigram language model as the original engine.

As a pre-research, we leverage a large corpus we have available from speech-synthesis work to study the benefit of using massive amounts of speaker-specific training data. The speech waveforms in the corpus are recorded through a close talking microphone in a quiet environment, with 16,000 Hz sampling rate and 16-bit PCM quantization.

The whole corpus contains approximate 13,000 utterances from a single female speaker. After removing the sentences with *out of vocabulary* (OOV) words, the number of utterances is reduced to 12,806. Since the corpus covers about 90% texts of the *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus* (TIMIT) [8], it can be further subdivided into training and test sets just like TIMIT does in respect that the subdivision of texts in TIMIT satisfies the following criteria:

1. The amount of overlap of text material in training and test sets are minimized.
2. All the phonemes are covered in the test material.

The core test set in TIMIT which is the minimum recommended set for test purposes contains 192 different texts, while 624 texts are reserved by the complete test set. In this paper, we take the subset whose text is the same with the TIMIT core test set as the evaluation test set, while we look upon the subset whose text is the same with the TIMIT complete test set but the core test set as the development test set. The development test set is used for model-parameter-tuning purposes. The rest of the speaker-specific corpus forms the training database. Thus, after removing the OOVs, the resulting data subdivision of the whole corpus is given in Table 1.

Table 1. Subdivision of training and test sets

Data Set		#Utterances	#Hours
training set		12,272	15.37
test sets	dev.	364	0.46
	eval.	170	0.21

To compare the performance of MSSR with that of SASR in the case of massive amounts of speaker-specific training data, we perform speaker-adaptation experiments based on the same training- and test-set subdivision with MSSR. In this study, both MLLR and MAP adaptation techniques are used. During the adaptation, statistics are accumulated for each Gaussian based on the recognition alignment. For every adaptation pass, MLLR, for both mean and variance, is applied first where given the amount of data, appropriate regression classes are selected and full transformation is computed and applied. MAP is subsequently applied for every Gaussian after MLLR. Once the number of adaptation frames exceeds a certain threshold (about 1.5 hours), the adaptation will stop applying MLLR and only apply MAP adaptation since at this point MAP is more effective than MLLR.

4. TRAINING-SUBSET SELECTION

In order to study the behavior of MSSR with different amounts of training data, 5 incremental training subsets are selected from the speaker-specific training corpus using the following criteria:

1. The first subset is the smallest data set covering all triphones seen in the training corpus. Different sentence-selection sequence will result in different content in this subset.
2. The last (largest) subset is actually the whole training corpus.
3. Each of other training subsets is then constructed by putting together the previous subset and some randomly selected new data. Thus, all subsets have the inclusion relation like S1 ⊂ S2 ⊂ S3 ⊂ S4 ⊂ S5. The difference number of utterances between every two adjacent subsets is almost the same.

In this paper, we repeat this training-subset-selection process 5 times to obtain 5 groups of training subsets, say G1–G5, with different sentence-selection order for the first subset and different random increment for the others. The distribution of the numbers of utterances and hours of every training subsets in all groups follows Table 2.

5. EXPERIMENTAL RESULTS

5.1. SISR result

The modified Microsoft English speech-recognition engine where whole-word models are discarded is evaluated on the evaluation test set described in Table 1. The word recognition error rate is

Table 2. Training-subset distribution

Training Subset	Group	#Utterances	#Hours
S1	G1	5,592	7.66
	G2	5,610	7.68
	G3	5,614	7.68
	G4	5,599	7.64
	G5	5,588	7.67
	average	**5,600**	**7.67**
S2	G1	7,262	9.59
	G2	7,275	9.62
	G3	7,278	9.62
	G4	7,267	9.58
	G5	7,259	9.61
	average	**7,268**	**9.60**
S3	G1	8,932	11.51
	G2	8,940	11.50
	G3	8,942	11.52
	G4	8,935	11.51
	G5	8,930	11.54
	average	**8,936**	**11.52**
S4	G1	10,602	13.45
	G2	10,605	13.43
	G3	10,606	13.45
	G4	10,603	13.45
	G5	10,601	13.45
	average	**10,603**	**13.45**
S5	G1–G5	12,272	15.37

14.58%. This SISR system is referred to as the baseline system in comparison with both SASR and MSSR.

5.2. SASR results

We perform speaker-adaptation tests using all 5 training subsets in each group described in Section 4. Performance of SASR using some smaller randomly-selected subsets are studied as well in order to obtain a complete view of SASR. The smaller training subsets in each group contain 5, 10, 20, 50, 100, 200, 500, 1000, 2500, and 4000 sentences, respectively.

In this experiment, the SISR model is used as the initial model and adapted to the specific speaker. The word recognition error rate of the SASR system trained on each training subsets shown in Table 2 is listed in Table 3. The average *relative error reduction* (RER) of each training subset to the SISR system is listed in the last row as well. The total average RER of all 21 training subsets is about 42.6%. Varying the number of training sentences from 5 to 4000 in each group, together with the numbers of sentences in the third column in Table 3, a curve of average WERs of 5 training subsets with SASR is plotted in Figure 1.

As shown in Figure 1, there is a deep drop of WER when 5 sentences are used in adaptation. Further increase of the number of adaptive sentences, less than a threshold, continues to lead to obvious improvement. This phenomenon is consistent with well-accepted theory within the community. From the results, we would also assume that the inflection point appears at about 1,000 sentences (1.3 hours). After more than 1,000 sentences being added, the SASR does not response with the same significant improvement as before.

Table 3. SASR results

Training Subset		S1	S2	S3	S4	S5
WER (%)	G1	9.12	8.69	8.41	8.33	7.95
	G2	8.69	8.41	8.41	7.97	
	G3	8.76	8.41	8.41	7.97	
	G4	8.62	8.55	8.48	8.48	
	G5	8.55	8.48	8.62	8.26	
	Avg	8.75	8.51	8.47	8.20	7.95
Avg RER vs. SISR (%)		40.0	41.6	41.9	43.8	45.5

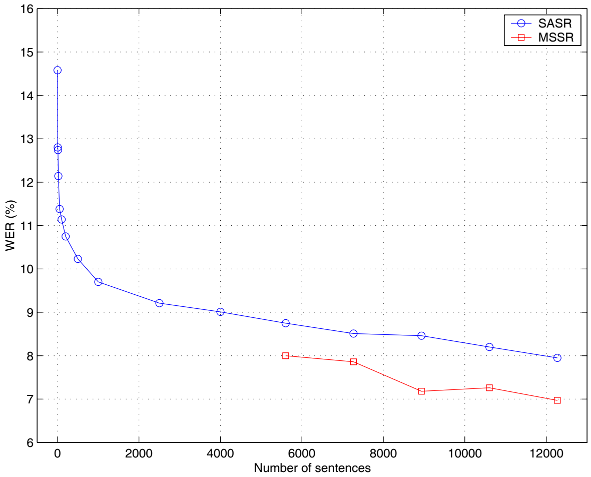

Fig. 1. Performance comparison between SASR and MSSR.

5.3. MSSR results

The acoustic models in MSSR are trained via a set of modified HTK tools which support SAPI's front end and can accurately control the number of senones surviving in state clustering. The training process is similar with that in [9].

Three kinds of parameters are tuned in this study. They are 1) the number of senones, n_s, 2) the number of mixtures, n_m, and 3) the minimum number of samples, n_d, for each senone in state-clustering step. When experimenting with different n_s and n_m, we found that tuning n_m is more effective. Therefore, we fix n_s to 2,750 and place emphasis on tuning n_m. For the third parameter, only 400 and 500 are investigated with $n_s = 2,750$ and finally $n_d = 500$ is selected.

We vary n_m from 8 to 48 in our paper since so far the decoding program cannot work well for more than 50 mixtures. Figure 2 displays the WERs of all 5 training subsets averaged on different groups versus the number of mixtures, n_m, which is evaluated on the development test set. Different amounts of training data may require different number of mixtures, as shown in Figure 2. From this figure, we would also conclude that within a certain range, the smaller the training subset is, the fewer the mixtures are needed. For example, the system trained on subset S1 has the best performance at 28 mixtures. Both up and down of number of mixtures do harm. But systems trained on subsets S4 and S5 need at least 48 mixtures. To evaluate the performance of MSSR on the evaluation test set as in SISR and SASR, for each training subset we adopt the best n_m as its system parameter. Table 4 lists the results.

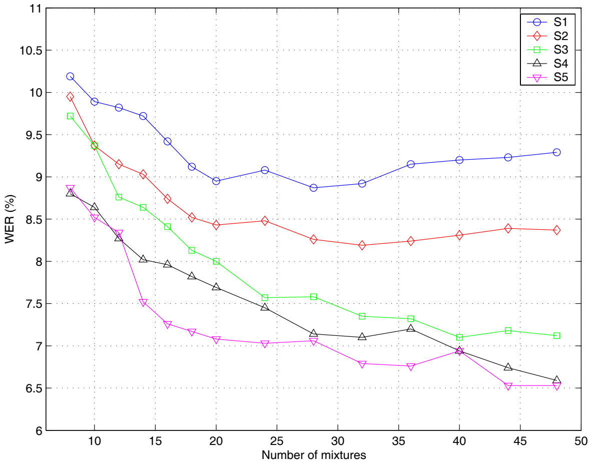

Fig. 2. Average WER of MSSR on development test set versus n_m.

Table 4. MSSR results

Subset		S1	S2	S3	S4	S5
n_m		28	32	40	48	48
WER (%)	G1	8.55	8.33	7.69	6.25	6.97
	G2	8.05	8.55	6.25	9.05	
	G3	7.90	6.68	7.11	6.61	
	G4	7.90	7.33	6.83	7.11	
	G5	7.62	8.41	8.05	7.26	
	Avg	8.00	7.86	7.19	7.26	6.97
Avg RER vs. (%)	SISR	45.1	46.1	50.7	50.2	52.2
	SASR	8.6	7.6	15.0	11.5	12.3

As shown in Table 4, for MSSR the WER is reduced by about 2 times in comparison with the SISR system. MSSR also performs better than SASR with an average RER of about 11% on massive amounts of training data. In order to compare MSSR with SASR, the average WERs in Table 4 are also displayed in Figure 1. From this figure, MSSR via less than 8 hours speech is already comparable with SASR via more than 15 hours speech. The performance of MSSR still has significant improvement even when the amount of training sentences exceeds 5,000.

6. SUMMARY AND FUTURE WORK

In this paper, we leverage the large corpus obtained from speech-synthesis work to study the benefit of using massive amounts of speaker-specific training data for LVCSR, i.e., MSSR. We demonstrate that by changing the focus from SISR to MSSR, word recognition error rate can drop very significantly. In comparison with SASR system, MSSR also performs better since model parameters can be tuned to be suitable to the particular individual. With 12,272 training sentences (more than 15 hours), the error reductions relative to SISR and SASR are 52.2% and 12.3%, respectively. Even with only 5,600 sentences (less than 8 hours), MSSR also outperforms both SISR and SASR by RERs 45.1% and 8.6%.

Future research work may includes replacing the reading-style training data in MSSR by spontaneous speech recorded in one's daily lives. Investigation of benefit of using speaker-dependent lexicon and language model would be another research direction.

7. ACKNOWLEDGMENT

The authors would like to thank the intern students Xiaowo Wang and Chang Hu from Tsinghua University for their assistant in doing the experiments.

8. REFERENCES

[1] K. F. Lee and X. D. Huang, "On Speaker-Independent, Speaker-Dependent, and Speaker-Adaptive Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, pp. 150-157, vol. 1, no. 2, April 1993.

[2] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A Compact Model for Speaker-Adaptive Training," *in Proc. ICSLP'96*, vol. 2, pp. 1137-1140, 1996.

[3] C. Huang, T. Chen and E. Chang, "Speaker Selection Training for Large Vocabulary Continuous Speech Recognition," *in Proc. ICASSP2002*, vol. 1, pp. 609-612, 2002.

[4] C. Huang, T. Chen and E. Chang, "Adaptive Model Combination for Dynamic Speaker Selection Training," *in Proc. ICSLP2002*, vol. 1, pp. 65-68, 2002.

[5] H. Yu, T. Tomokiyo, Z. Wang, and A. Waibel, "New Developments in Automatic Meeting Transcription," *in Proc. ICSLP2000*, 2000.

[6] K.-T. Chen, S.-L. Chuang, F. Seide, H.-M. Wang, L.-F. Chien, and E. Chang, "New Word Learning for Spoken Document Processing Through Discovery of Comparable Texts from External Resources," *in Proc. MSDR2003*, pp. 79-84, 2003.

[7] J. L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, 37(1-2):89-108, 2002.

[8] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

[9] E. Chang, Y. Shi, J. L. Zhou, and C. Huang, "Speech Lab in a Box: A Mandarin Speech Toolbox to Jumpstart Speech Related Research," *in Proc. Eurospeech2001*, pp. 2799-2803, 2001.