USING MULTIPLE VERSIONS OF SPEECH INPUT IN PHONE RECOGNITION

Mark Liberman[1], Jiahong Yuan[1], Andreas Stolcke[2], Wen Wang[3], Vikramjit Mitra[3]

[1]University of Pennsylvania, [2]Microsoft Research, [3]SRI International

ABSTRACT

This study investigates the use of multiple versions of the same speech unit in automatic phone recognition. Two methods were applied to combine multiple utterance versions in decoding: cross forced-alignment and *n*-best ROVER. The phone error rate was reduced from 15% to 2% on isolated words and from 33% to 19% on TIMIT sentences. The error rate was reduced the most when the second version was added, and less so as each additional version was added. Depending on the language model weight, it might be better to use the language model only in *n*-best generation, but omit it in scoring the hypotheses applied to the combination methods. *N*-best ROVER effectiveness may be enhanced by lowering the language model weight.

Index Terms— Forced alignment, *N*-best ROVER, phone recognition, multiple utterance versions

1. INTRODUCTION

The performance of the state-of-the-art Hidden Markov Model based automatic speech recognition (ASR) systems has improved substantially over the last several decades. To further reduce error rates for ASR is a challenging problem. In this study, we investigate the benefit of using multiple versions of a speech unit in automatic phone recognition.

It is not uncommon for human listeners to have the benefit of multiple versions of the same utterance. For example, children may rely on multiple examples of the same word to acquire a lexicon, and speakers often ask for confirmation or repetition in conversation. However, most ASR systems recognize individual utterances independently, and the problem of simultaneously decoding multiple versions of the same speech unit has not been widely addressed. Several studies have attempted to deal with this problem. Haeb-Umbach et al. [1] applied both multiple-candidate transcription method and average transcription method in automatic transcription of unknown words. In the first method, multiple transcriptions of an unknown word, one for each utterance version of the word, were compared to all utterance versions of the word, and the obtained transcription was the one for which the product of the likelihoods of all utterance versions given the transcription was maximum. In the second method, an "average utterance" was generated by training a whole-word model from all the word's utterance versions, and the obtained transcription was the one that had the highest likelihood on this average utterance. Wu and Gupta [2] proposed a word-network-based algorithm for simultaneous decoding of multiple utterance versions. The algorithm merged independently scored lattices, one for each utterance version, to form a lattice scored jointly from all utterance versions, and then found an optimal path through the combined lattice. Nair and Sreenivas [3-4] applied Multi Pattern Dynamic Time Warping to determine the optimal path in the joint space of all utterance versions, and applied Multi Pattern Joint Likelihood algorithms to determine the best state sequence for all utterance versions jointly. Related work has also been done in pronunciation modeling [5-6] and decoding partly repeated utterances in voice search [7-8].

In this study, we investigate how the error rate is reduced when more utterance versions are available in automatic phone recognition for both isolated words and sentences. We apply two popular techniques to combine multiple utterance versions in decoding - forced alignment and ROVER. In the following sections, the methods are first described in Section 2, followed by the experiments on isolated words and sentences in Section 3 and 4 respectively. Finally, Section 5 completes the paper with conclusions and a discussion of future research issues.

2. CROSS FORCED-ALIGNMENT AND N-BEST ROVER

Different utterance versions of the same word or sentence, whether from different speakers or from repetitions of the same speaker, have different acoustic patterns. Some versions may be more accurately recognized than others by an ASR system. Suppose there are m utterance versions for a speech unit and each version has a set of n-best hypotheses generated by an ASR system: $H(U_i) = \{h_{i1}, h_{i2}, \ldots, h_{in}\}$, $i = 1, 2, .., m$; we can pool the hypotheses of all utterance versions, $H_{all} = \bigcup H(U_i) = \{h_{11}, h_{12}, \ldots, h_{mn}\}$, then apply forced alignment between each hypothesis and each utterance version respectively. Therefore, every hypothesis has m recognizer scores, one from alignment with each of the utterance versions. The optimal hypothesis h^* has the best recognizer scores with respect to a statistic of the m scores. If the mean score is used, for example, we have:

$$h^* = \underset{h \in H_{all}}{\operatorname{argmax}} \frac{\sum_{i=1}^{m} \log P(U_i|h)}{m}$$

We call this method "cross forced-alignment". Cross forced-alignment selects a hypothesis from the independently generated *n*-best lists that jointly maximizes the likelihood of all utterance versions.

As an utterance gets longer, it becomes increasingly unlikely that any single hypothesis in the *n*-best lists accurately transcribes all of its phones. However, different portions of different utterance versions may still be correctly recognized. These portions can be combined to make an optimal hypothesis using a combination technique such as ROVER [9]. ROVER was developed to combine the 1-best outputs from multiple ASR systems to produce a composite output that has a lower error rate. It consists of two steps. First, the outputs are aligned to build a word transition network; secondly, the resulting network is searched and the best scoring word at each node is selected. Stolcke et al. [10] extended ROVER to *n*-best lists from multiple systems. Each system yields a posterior probability estimate at the token (word or phone) level, and these multiple estimates are combined in a weighted fashion. Finally, the token with the highest posterior probability at each position is chosen, by which we minimize the expected token-level error rate of the hypothesis.

In this study, we applied *n*-best ROVER, implemented in the SRILM toolkit [11], to *n*-best lists generated by the same system for multiple utterance versions. In this *n*-best ROVER approach, word posterior estimates used in the voting process were computed independently for each utterance version, i.e., from the *n*-best list of each utterance version respectively. We also combined cross forced-alignment with *n*-best ROVER, in which the recognizer score of every hypothesis in the *n*-best list pool was obtained based on the recognizer scores of forced alignment between the hypothesis and all utterance versions, and the *n*-best lists with these new cross forced-alignment scores were used for ROVER.

3. SPEECH CONTROLLED COMPUTING EXPERIMENTS

The Speech Controlled Computing corpus [12] consists of the recordings of 125 speakers of American English from four dialect regions, three age groups and two gender groups. Each speaker read a randomized word list consisting of 2,100 word items - 100 distinct words appearing 21 times each. Randomly selected 100 versions for each of the 100 words from all speakers, totally 10,000 utterances, were used for testing, and the rest for training. Monophone HMM and GMM acoustic models, with the standard 39 MFCC features and 256 Gaussian mixtures, were trained using the CMU pronouncing dictionary [13] and the HTK toolkit [14]. The acoustic models had a word accuracy of more than 99.9% on the test set.

In the following experiment of automatic phone recognition, a phone network in which all phones may appear at any position with equal probability was used as the language model. We assumed all versions of a word have the same phone sequence, which was used as the "gold standard" in the tests. The 10,000 test utterances were divided into 10 test sets. Each set contained 10 versions for each word, totally 1,000 utterances. Different numbers of utterance versions were used in decoding. When only one utterance version for each word was used, it was a typical automatic phone recognition task. In this case, we conducted 10 tests for every test set, each using one of the 10 utterance versions, and there were 100 tests in total for the 10 sets. When two or more utterance versions were used, the independently generated 5-best lists of the utterance versions were pooled. Each hypothesis in the pool was forced aligned with all utterance versions respectively, and the hypothesis with the highest average log probability score from forced alignment was selected. There were 45 tests for each test set when two utterance versions were used, i.e., 45 combinations of choosing two versions from 10; 120 tests for each test set when three utterance versions were used, etc. Table 1 and Figure 1 (on next page) show the phone error rates when using different numbers of utterance versions.

Table 1. Averaged phone error rates on isolated words when using different numbers of utterance versions for each word.

Number of versions	Number of tests	Mean phone error rate	Relative reduction from using one fewer version
1	100	14.89%	-
2	450	6.18%	58.5%
3	1200	3.77%	39.0%
4	2100	2.85%	24.4%
5	2520	2.41%	15.4%
6	2100	2.15%	10.8%
7	1200	2.01%	6.5%
8	450	1.93%	4.0%
9	100	1.89%	2.1%
10	10	1.90%	-0.5%

With the same acoustic models, the phone error rate was reduced from approximately 15% when only one utterance version was used to less than 2% when eight or more versions were used. The error rate was reduced by nearly 60% when the second version was added, and by increasingly smaller additional factors as more versions were added, reaching a minimum after seven or eight versions. At this stage, the most frequent errors were confusions between pairs such as /r/ and /er0/, /iy0/ and /iy1/, etc. Those "errors" may result from pronunciation variability in the data, because we assumed no alternative pronunciations in the experiment.

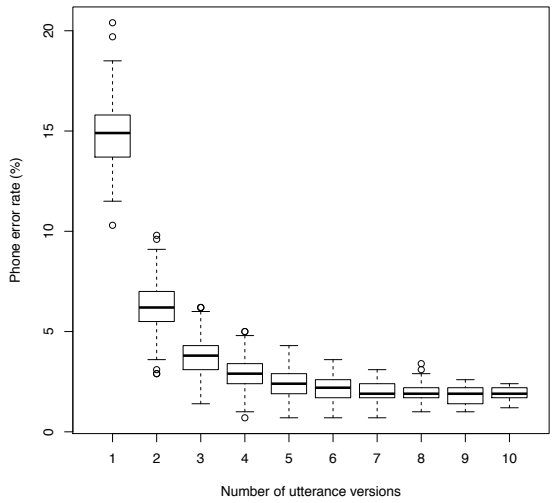

Figure 1. Boxplots of phone error rates on isolated words when using different numbers of utterance versions.

4. TIMIT EXPERIMENTS

The SX sentences in the TIMIT corpus [15] were each spoken by seven speakers. 150 SX sentences, each having seven utterance versions, were randomly selected for testing. The 150 test sentences were divided into five test sets, each having 210 copies of 30 sentences. All other sentences excluding the SA sentences, 3,990 utterances in total, were used for training the acoustic models. Triphone HMM and GMM acoustic models, with the standard 39 MFCC features, 16 Gaussian mixtures and 1,109 tied states, were trained using the TIMIT word transcripts, the TIMIT pronouncing dictionary, and the HTK toolkit. Every word in the TIMIT pronouncing dictionary has only one phone sequence, which was used for both training and scoring. The TIMIT phone transcriptions were not used. A phone bigram language model was trained on the phone sequences of the training utterances, and was used as the language model in the following experiment of automatic phone recognition. Following Lee and Hon [16], the following phone pairs were merged for scoring: (/ix/, /ih/), (/ax/, /ah/), (/ao/, /aa/), (/zh/, /sh/).

We first applied cross forced-alignment to the test sets. As in the experiment using the Speech Controlled Computing corpus, different numbers of utterance versions were used in decoding. When two or more utterance versions were used, the independently generated 100-best lists of the utterance versions were pooled, and the hypothesis with the highest average log probability score from cross forced-alignment was selected. The grammar scale factor used for generating 100-best lists was set at 10. Table 2 and Figure 2 show the phone error rates when using different numbers of utterance versions.

Table 2. Averaged phone error rates on TIMIT when using different numbers of utterance versions for each sentence.

Number of versions	Number of tests	Mean phone error rate	Relative reduction from using one fewer version
1	35	34.27%	-
2	105	28.36%	17.2%
3	175	25.44%	10.3%
4	175	23.69%	6.9%
5	105	22.59%	4.6%
6	35	21.86%	3.2%
7	5	21.06%	3.7%

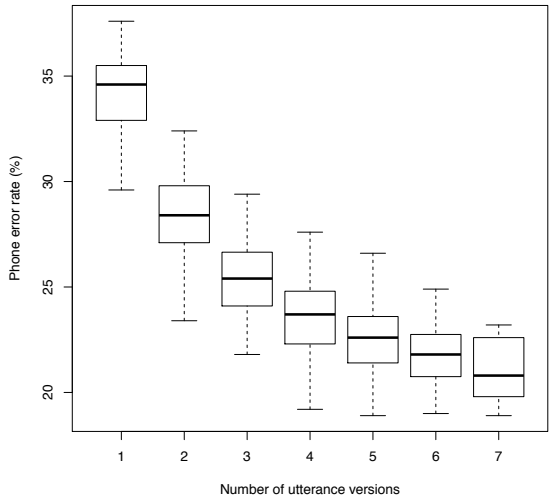

Figure 2. Boxplots of phone error rates on TIMIT when using different numbers of utterance versions.

With the same acoustic models and language model, the phone error rate was reduced from 34% when only one utterance version was used, to 21% when seven versions were used. The error reduction from using more utterance versions on TIMIT showed a pattern similar to that on the Speech Controlled Computing corpus. Generally, the error rate was reduced the most when the second version was added, and less so when additional versions were added. However, the error rate was reduced more gradually on TIMIT than on the Speech Controlled Computing corpus, and the overall error reduction was also less substantial on TIMIT. We hypothesize that this is because cross forced-alignment only selects complete utterance-level hypotheses from *n*-best lists, but does not combine hypotheses. When utterances are longer, such as those of TIMIT, no hypothesis in the pool may be perfect, and one hypothesis may be more accurate than another only in some portions. In this case, *n*-best ROVER should be able to achieve greater error reduction.

N-best ROVER relies on recognizer scores to compute utterance-level posterior probabilities (which are then used

to derive word-level posteriors). When a language model is used to generate *n*-best lists, the recognizer scores include both acoustic and language model scores. It is not obvious whether we should use the acoustic scores only or the overall recognizer scores for *n*-best ROVER, given that the hypotheses were generated with a language model and the goal is to concatenate portions of different hypotheses to make an optimal hypothesis for all utterance versions of the same sentence. It is also not obvious how the weight of the language model with respect to the acoustic model may affect *n*-best ROVER. On one hand, a more constraining language model may provide overall more accurate hypotheses in the *n*-best lists; on the other hand, a less constraining language model may provide more variability between the *n*-best lists. Both more accurate and more variable hypotheses may benefit *n*-best ROVER. In the following experiment, we used three grammar scale factors, 1.0, 5.0 and 10.0, which determine the weight of the language model relative to the acoustic model, and used both acoustic scores only and acoustic plus language model scores as *n*-best scores. We also ran a test without any language model; instead, a phone network in which all phones may appear at any position with equal probability was used as the language model. *N*-best ROVER with and without cross forced-alignment were applied. Cross forced-alignment was also applied by itself for comparison. The recognizer scores from forced alignment were acoustic scores only; the language model scores for forced alignment, when used, were computed separately. All seven utterance versions were used in decoding, 100-best lists were generated for each utterance version. The overall phone error rates for the 150 test sentences from different methods are listed in Table 3.

Table 3. Phone error rates on TIMIT when using seven utterance versions and different combination methods.

	No language model	s = 1.0	s = 5.0	s = 10.0
Baseline	44.8%	40.2%	34.0%	**33.4%**
F + A	30.7%	26.8%	22.6%	**21.0%**
F + AL		26.2%	**21.8%**	22.4%
R + A	22.9%	20.5%	**19.6%**	21.1%
R + AL		20.6%	**19.2%**	21.5%
FR + A	22.2%	20.0%	**18.9%**	20.3%
FR + AL		20.0%	**18.8%**	21.1%

Baseline: the seven versions were independently decoded. **F**: Cross forced-alignment. **R**: *N*-best ROVER. **FR**: *N*-best ROVER with cross forced-alignment. **A**: Acoustic scores only. **AL**: Acoustic plus language model scores (**AL** and **A** are the same when no language model was used). **s**: grammar scale factor; it is used both for generating *n*-best lists and for weighting the language model scores, when used, in cross forced-alignment and *n*-best ROVER.

From Table 3 we can see that both cross forced-alignment and *n*-best ROVER significantly reduced the phone error rate compared to the baseline system, for which multiple versions of a sentence were independently decoded. *N*-best ROVER performed better than cross forced-alignment, more so when the weight of the language model was smaller or there was no constraining language model. The optimal weight of the language model for the baseline system was different from that for cross forced-alignment and *n*-best ROVER. *N*-best ROVER preferred a smaller language model weight than the baseline system, presumably because that increases the independence of the generated hypotheses. *N*-best ROVER with cross forced-alignment performed only slightly better than *n*-best ROVER alone. Depending on the language model weight, it might be better to use the language model only in *n*-best generation, but omit it in scoring the hypotheses. For example, using acoustic scores only performed better than using acoustic plus language model scores for all the combination methods when the language model weight was 10.0: Cross forced-alignment – 21.0 vs. 22.4%; *N*-best ROVER – 21.1% vs. 21.5%; *N*-best ROVER with cross forced-alignment – 20.3% vs. 21.1%.

5. CONCLUSIONS

We have demonstrated the benefit of phone-level decoding of multiple versions of a speech unit, whether it is a single word or a sentence. Using cross forced-alignment and *n*-best ROVER to combine multiple utterance versions in decoding, the phone error rate was reduced from 15% to 2% on isolated words and from 33% to 19% on TIMIT sentences. The error rate was reduced the most when the second version was added, and was reduced by successively smaller differences as additional versions were added. When a phonotactic language model was used, the optimal weight of the language model for *n*-best ROVER was smaller than that for the baseline system in which multiple utterance versions were decoded independently. Whether language model scores helped in applying cross forced-alignment and *n*-best ROVER depends on the weight of the language model.

In this study, we assumed all utterance versions of the same speech unit have the same phone sequences. In future work, it will be desirable to generalize to alternative pronunciations or multiple utterance versions with speech errors or disfluencies. We used the mean of the recognizer scores to determine the optimal hypothesis in cross forced-alignment. Other statistics, such as the maximum, the minimum, and the variance, and how to use of a language model in decoding multiple utterance versions should also be investigated in future studies.

6. ACKNOWLEDGMENTS

This work is supported in part by NSF grant IIS-0964556.

7. REFERENCES

[1] Haeb-Umbach, R., Beyerlein, P., Thelen, E., "Automatic transcription of unknown words in a speech recognition system," *Proceedings of ICASSP 1995*, pp. 840–843, 1995.

[2] Wu, J., Gupta, V., "Application of simultaneous decoding algorithms to automatic transcription of known and unknown words," *Proceedings of ICASSP 1999,* pp. 589–592, 1999.

[3] Nair, N.U., Sreenivas, T.V., "Joint decoding of multiple speech patterns for robust speech recognition," *Proceedings of ASRU 2007*, pp. 93–98, 2007.

[4] Nair, N.U., Sreenivas, T.V., "Joint evaluation of multiple speech patterns for speech recognition and training," *Computer Speech and Language*, 24, pp. 307-340, 2010.

[5] Holter, T., Svendsen, T., "Maximum likelihood modeling of pronunciation variation," *Proceedings of ESCA Workshop on Modeling Pronunciation Variation for ASR*, pp. 63–66, 1998.

[6] Singh, R., Raj, B., Stern, R.M., "Automatic generation of subword units for speech recognition systems," *IEEE Trans. Speech Audio Proc.* 10, pp. 89–99, 2002.

[7] Bohus, D., Zweig, G., Nguyen, P., Li, X., " Joint N-best rescoring for repeated utterances in spoken dialog systems," *Proceedings of SLT 2008*, pp. 133-146, 2008.

[8] Zweig, G., Bohus, D., Li, X., Nguyen, P., "Structured models for joint decoding of repeated utterances," *Proceedings of Interspeech 2008*, pp. 1157-1160, 2008.

[9] Fiscus, J.G., "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," *Proceedings of ASRU 1997*, pp. 347-354, 1997.

[10] Stolcke, A., Bratt, H., Butzberger, J., Franco, H., Rao Gadde, V. R., Plauche, M., Richey, C., Shriberg, E., Sonmez, K., Weng, F., Zheng J., "The SRI March 2000 Hub-5 Conversational Speech Transcription System," *Proceedings NIST Speech Transcription Workshop, 2000*.

[11] Stolcke, A., SRILM - An Extensible Language Modeling Toolkit, *Proceedings of ICSLP 2002*, pp. 901-904, 2002. http://www.speech.sri.com/projects/srilm/.

[12] Cieri, C., et al., *Speech Controlled Computing* (LDC2006S30), Linguistic Data Consortium, 2006.

[13] The CMU Pronouncing dictionary: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[14] The Hidden Markov Model Toolkit (HTK): http://htk.eng.cam.ac.uk/

[15] Garofolo, J.S., *TIMIT Acoustic-Phonetic Continuous Speech Corpus* (LDC93S1), Linguistic Data Consortium, 1993.

[16] Lee, K.F., Hon, H., "Speaker-independent phone recognition using Hidden Markov models," *IEEE Trans. Acoustics Speech Signal Proc.*, 37, pp. 1641-1648, 1989.