A Comparison of Speech Recognizers Created Using Manually-Aligned and Automatically-Aligned Training Data

John-Paul Hosom

Center for Spoken Language Understanding (CSLU)
Oregon Graduate Institute of Science and Technology (OGI)
P.O. Box 91000, Portland Oregon 97291-1000 USA
e-mail: hosom@cse.ogi.edu   www: http://www.cse.ogi.edu/CSLU

ABSTRACT

It is generally acknowledged that time-alignment of phonemes is of better quality when obtained from manual segmentation as compared to automatic segmentation. However, empirical evidence for this belief is sparse. This paper describes a controlled study of two recognizers created using manually-aligned and automatically-aligned training data. Both recognizers were trained on the digits task using telephone-channel continuous speech. The manual alignments were generated by expert labelers, while the automatic alignments were obtained from our best general-purpose forced alignment system. The two recognizers were trained and evaluated using the same set of data and parameters whenever possible. The results of the recognizer trained on manually-aligned data were 97.54% word accuracy and 90.18% sentence accuracy. The results for the recognizer trained on automatically-aligned data were 97.24% word accuracy and 88.80% sentence accuracy. This represents an 11% reduction in error at the word level and a 12% reduction in error at the sentence level. The sentence-level results are statistically significant using McNemar's test, with p=0.002.

1. INTRODUCTION

The phonetic alignment of speech is thought to be performed better by humans than by computers. Evidence for this belief is seen in the statements of researchers in the field. Andrej Ljolje notes that "due to the … inherent limits in the parameterization of the speech signal and the speech model structure, the accuracy of the transcription [by automatic methods] is inferior to that achieved by human transcribers." [1]. Piero Cosi states that "The accuracy of automatic alignment systems will always be checked using references manually segmented by phonetic or speech communication experts." [2] Stephen Cox reports that "It is well known that … variation [of manual alignments] is generally small when compared with alignments produced by automatic systems." [3]

Empirical evidence for this prevalent belief is scarce, however. This paper reports the results of a controlled study of the performance of two recognizers; one trained with manual alignments, and the other trained with automatic alignments. Both recognizers were trained using the CSLU Toolkit, on the OGI Numbers corpus.

The recognizers in the CSLU Toolkit use a hybrid HMM/ANN framework [4]. In these systems, frame-based recognition is performed using context-dependent sub-phonetic states, where the state probability estimation is computed using a neural network.

We have developed a set of procedures within the Toolkit for training special-purpose recognizers for tasks such as continuous digit recognition. This method is simple enough that a bright high-school student can complete the tutorial in a few days. On the continuous digits task, the training procedure yields recognition results that compare favorably to standard HMM systems [4].

2. CORPUS

The OGI 30K Numbers corpus [5] was used for training, development, and testing. The data in this corpus were collected from thousands of people within the United States who recited their telephone number, street address, zip code, or other numeric information over the telephone in a natural speaking style. Because the data were collected from a large number of speakers from different backgrounds in different environments, the corpus contains a noticeable amount of breath noise, glottalization, background noise (including music), and other "real-life" complications. Of almost 15,000 utterances, approximately 6600 utterances have been transcribed and time-aligned at the phonetic level by professional labelers. For the experiments reported here, we used only those utterances that consist entirely of digits (zero through nine and "oh"). Before separating the data into training, development, and test sets, about 5% of the corpus was culled for independent testing and set aside. Three speaker-independent partitions were created from the remaining data: 3/5 for training (6087 files, of which 2547 were hand-labeled), 1/5 for development (2110 files), and 1/5 for testing (2169 files). The development partition was further split into five sets, and the development results reported in this paper are for the first of these five sets (423 files).

3. MANUAL-ALIGNMENT SYSTEM

The manual alignment system, referred to here as the System M, was created in two passes. The first past trained a recognizer using all available manually-aligned data in the training set. The second pass used the recognizer created in the first pass to automatically re-align the data and train a new system. This second pass often yields better performance because the automatic alignment is capable of specifying the alignments of sub-phonetic categories, which is not possible with manual alignments.

For training System M, hand-labeled phonetic symbols are mapped, if necessary, to a consistent set of symbols for each word, /oU 9r/ (in "four") is merged into one />r/ phone, and /kh s/ (in "six") is merged into one /ks/ phone. (Phonetic symbols are written in Worldbet).

The system is trained to recognize context-dependent units. For left and right contexts, pauses and stop closures are mapped to the symbol /uc/ (unvoiced closure), and dentals (/th/, /s/, and the right half of /ks/) are mapped to the broad-category symbol /den/; otherwise the contexts are phoneme-specific. Each phoneme can be split into one, two, or three parts. The left part is dependent on the context of the preceding phoneme (or phonetic broad category), the center part (if any) is context independent, and the right part is dependent on the following phoneme (or phonetic broad category). Phonemes that remain as a one-part phoneme can either be context-independent or be dependent on the following phoneme.

The system is trained using 13 MFCC features (12 cepstral coefficients and 1 energy parameter) plus their delta values, with a 10-msec frame rate. The input to the network consists of the features for the frame to be classified, as well as the features for frames at -60, -30, 30, and 60 msec relative to the frame to be classified (for a total of 130 input values). As many as 2000 samples per category are collected for training. Neural-network training is done with standard back-propagation on a fully-connected feed-forward network. The training is adjusted to use the negative penalty modification proposed by Wei and van Vuuren [6]. With this method, the non-uniform distribution of context-dependent classes that is dependent on the order of words in the training database is compensated for by flattening the class priors of infrequently occurring classes; this compensation allows better modeling for an utterance in which the order of the words can not be predicted.

During the Viterbi search, transition probabilities are set to be all equally likely, so that no assumptions are made about the likelihood of one category following another category. The search was constrained to minimize insertion errors by having minimum duration values for each category, where the minimum value for a category was computed as the value at two standard deviations from the mean duration. During the search, category durations less than the minimum value are penalized by a value proportional to the difference between the minimum duration and the proposed duration.

The grammar allows any number of digits in any order, with an optional silence between digits. In addition, a "garbage" word is allowed at the beginning and end of each utterance to account for sounds not in the vocabulary. The "garbage" word is defined as a word with a single context-independent category; the value of this category is not an output of the neural network, but is computed as the N^{th}-highest output from the neural network at each frame [7]. In this study, N was set to 5.

Training is done for 30 iterations, and the "best" network iteration is determined by word-level evaluation of each iteration on the development set data. This "best" network is then used to force-align the same training utterances, and training and evaluation are repeated to determine the final digits network for System M.

4. AUTOMATIC-ALIGNMENT SYSTEM

The automatic-alignment system, referred to here as System A, was also trained in two passes. In the first pass, label alignments were created using the CSLU Toolkit's general-purpose recognizer, and network training was performed on these labels. In the second pass, the recognizer created in the first pass was used to automatically re-align the data and train a new recognizer, just as was done for System M.

The automatic alignment system in the CSLU Toolkit is a forced-alignment system. Forced alignment uses an existing HMM (in this case, an HMM/ANN) recognizer and constrains the search path to be the correct (known) sequence of phonemes. The search result provides not only the (known) correct answer, but the time locations of each phoneme. The forced alignment system in the Toolkit was trained on telephone-channel speech; the corpora used were the OGI Stories, Names, and Numbers corpora. The fact that the general-purpose system was trained on data from the Numbers corpus does perhaps give this system an advantage over other general-purpose forced alignment systems, and so the results reported here for the automatic alignment system may be higher than results generated with other systems.

The same training files, mapping, feature set, and negative-penalty modifications that were used in the creation of System M were used in the creation of System A. In addition, the same context-dependent units were specified. However, because the two alignment procedures yielded slightly different results, the number of data samples in each context-dependent category for System M is slightly different than the number of data samples in each context-dependent category for System A.

5. SYSTEM PARAMETERS

In this section, we will provide information specific to the training of each recognizer. Both recognizers were trained on 2547 files from the OGI Numbers corpus in the first pass, 6087 files in the second pass, developed using 2110 files, and evaluated for the final test using 2169 files. Each of these partitions of the Numbers corpus are speaker-independent.

Both recognizers used 130 input features, 200 nodes in the hidden layer, and 218 context-dependent output categories. Categories representing silence were context-independent, all consonants (except for /th/) and reduced vowels were split into two context-dependent parts, all long vowels and diphthongs were split into a context-dependent left third, a context-independent middle region, and a context-dependent right third. The /th/ phoneme was dependent on its right context only.

The first pass of each recognizer was trained using as many as 2000 samples per category, and the second pass of each recognizer was trained using all available data samples. The first pass of each recognizer was trained for 45 iterations, and the second pass of each recognizer was trained for 30 iterations.

6. RESULTS

The results of test-set evaluation on the first and second passes are summarized in Table 1. The 90.18% sentence-level result on 2169 files (sentences) for System M is significantly better than the 88.80% sentence-level result for System A, as evaluated using McNemar's test (*p*=0.002). The recognizer trained on manually-aligned data has an 11% relative reduction in error at the word level and a 12% relative reduction in error at the sentence level.

System	Word Accuracy	Sentence Accuracy	Reduction in Error
Automatic	97.24%	88.80%	n/a
Manual	97.54%	90.18%	11% (w), 12%(s)

Table 1: Test-set results for the recognizers trained on manually-aligned and automatically-aligned data. Evaluation was done on 2169 telephone-channel continuous-speech digit utterances (12437 words).

7. DISCUSSION

The 11% to 12% reduction in error obtained from these experiments indicates that phonetic alignments obtained from expert human labelers are, in fact, superior to alignments obtained from an automatic-alignment system. Furthermore, this superiority is reflected in the performance of recognizers trained on the phonetically-aligned data. These results are statistically significant, with *p*=0.002, indicating that the improvement in results with the manually-aligned data did not occur by chance.

8. ACKNOWLEDGEMENTS

The author would like to thank Ron Cole, Jacques de Villiers, Johan Schalkwyk, Kal Shobaki, and the CSLU member companies. This work was sponsored in part by the National Science Foundation (grant numbers GER-9354959 and IRI-9614217); the views expressed in this paper do not necessarily represent the views of the NSF.

REFERENCES

1. Ljolje, A., Hirschberg, J., and van Santen, J.P.H., "Automatic Speech Segmentation for Concatenative Inventory Selection." in *Progress in Speech Synthesis*, J.P.H. van Santen, R.W. Sproat, J. Olive, and J. Hirschberg, eds., Springer-Verlag, New York, 1997.
2. Cosi, P., Falavigna, D., and Omologo, M., "A Preliminary Statistical Evaluation of Manual and Automatic Segmentation Discrepancies", EuroSpeech '93, Genova, Italy, September 1991, pp. 693-696.
3. Cox, S., Brady, R., and Jackson, P., "Techniques for Accurate Automatic Annotation of Speech Waveforms", ICSLP'98, Sydney, Australia, December 1998, pp. 1947-1950.
4. Cosi, P., Hosom, J.P., Shalkwyk, J., Sutton, S., and Cole, R.A., "Connected Digit Recognition Experiments with the OGI Toolkit's Neural Network and HMM Based Recognizers," IVTTA-ETWR-98, Turin, Sep. 1998.
5. Cole, R.A., Fanty, M., Noel, M., and Lander, T., "Telephone Speech Corpus Development at CSLU," ICSLP-94, Yokohama, September 1994, pp. 1815-1818.
6. Wei, W. and Van Vuuren, S., "Improved Neural Network Training of Inter-Word Context Units for Connected Digit Recognition," ICASSP-98, vol. 1, Seattle, May 1998, pp. 497-500.
7. Boite, J.M., Bourlard, H., D'hoore, B., and Haesen, M., "A New Approach Towards Keyword Spotting," EUROSPEECH '93, Berlin, Sep. 1993, pp. 1273-1276.