

A DIPHONE-BASED DIGIT RECOGNITION SYSTEM USING NEURAL NETWORKS

John-Paul Hosom

Ronald A. Cole

Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology
P.O. Box 91000, Portland, Oregon 97291-1000 USA
{hosom, cole}@cse.ogi.edu <http://www.cse.ogi.edu/CSLU/>

ABSTRACT

In exploring new ways of looking at speech data, we have developed an alternative method of segmentation for training a neural-network-based digit-recognition system. Whereas previous methods segment the data into monophones, biphones, or triphones and train on each sub-phone unit in several broad-category contexts, our new method uses modified diphones to train on the regions of greatest spectral change as well as the regions of greatest stability. Although we account for regions of spectral stability, we do not *require* their presence in our word models. Empirical evidence for the advantage of this new method is seen by the 13% reduction in word-level error that was achieved on a test set of the OGI Numbers corpus. Comparison was made to a baseline system that used context-independent monophones and context-dependent biphones and triphones.

1. INTRODUCTION

Previous methods for training neural-network recognizers divide each phone for training into one, two, or three parts (monophones, biphones, or triphones). Training is then done on each phone or sub-phone, either as context-independent units, or in various broad-category contexts [1]. Context-dependent modeling is done in order to account for the coarticulatory effects of neighboring phones.

Context-dependent modeling, however, may not always provide appropriate information for distinguishing variations of a phone in different contexts. For example, in the task of continuous digit recognition, an /f/ may have the same coarticulatory effects on an /oU/ as silence would, leading to confusion between the words “oh” and “four.” A second difficulty in the current method of segmentation is that the number of subdivisions per phone is fixed. For phones that have a great variation in duration, dividing the phone into a fixed number of parts may not always yield satisfactory results. For example, the /I/ in the word “zero” can vary greatly in duration. (In the OGI Numbers cor-

pus [2], human labelers have found the duration of this phone to range from 20 to 280 msec.) If /I/ is modeled as a biphone, then long instances of /I/ may be poorly recognized during the middle section. Conversely, if /I/ is modeled as a triphone but is short in duration, then the required middle part of /I/ may have a low neural-network output score.

We propose a new segmentation method that offers the potential to overcome these two difficulties. In this new method, the diphone is the basis for segmentation, but an optional middle part is used for long phones. The word models used in recognition are also constructed to recognize diphone representations of the words, as well as accept an optional middle part for most phones. This diphone-based system classifies the regions of the speech signal at which there is the greatest change in the spectrum: the regions of transition between phones. By doing segmentation in this way, we hope to maximize the spectral differences between categories and thus improve the ability of the neural network to perform classification.

2. TASK

For investigating the performance of our diphone-based system, we selected the task of continuous-digit recognition of telephone-band speech.

2.1. Corpus

The corpus that we used for training, development, and testing was a subset of the OGI Numbers corpus [2] that contains only digits. This digits corpus contains many thousand utterances of digit strings spoken by a large number of people under various conditions. As a result, many aspects of “real-life” speech are present in the data, including noise, widely-varying energy levels, and dialect differences. We allocated three-fifths of the available data for training, and one-fifth each for development and testing.

2.2. Measurements

We measure the performance of both the baseline and the proposed method using the scoring algo-

rithm developed by NIST [3]. We report the overall accuracy as well as the percentage of substitutions, insertions, and deletions.

3. THE BASELINE RECOGNITION SYSTEM

In this section, we describe the method used to construct the baseline system. This system uses a neural network to classify fixed-length frames of speech into context-dependent phone-based categories. Construction of both the baseline system and the diphone system was done using the *cslush* speech-processing software package developed at OGI [4].

3.1. Baseline Segmentation Method

The context-dependent categories used in the baseline system are determined from speech data that have been hand-labeled at the phone level. Each phone is assigned a fixed number of parts to be split into, and a broad category of speech to which it belongs. If the phone is to be split into two parts, then one category is created for the left half of the phone in the context of the preceding phone's broad category. Another category is created for the right half of the phone in the context of the following phone's broad category. If the phone is to be split into three parts, then one category is created for the left third of the phone in the preceding phone's context, a second category is created for the middle third of the phone with no context, and a third category is created for the right third of the phone in the following phone's context. A monophone has one category that is not dependent on the contexts of surrounding phones.

A segmentation of the isolated word “two” using this method is illustrated in Figure 1. The phone /th/ is a two-part phone with the broad category “obstruent” (\$obs). The phone /u/ is a three-part phone. For the left third of /u/, the broad category “front-vowel” (\$fnt) is used because the /u/ follows a dental sound; the right third of /u/ has the broad category “back-vowel” (\$bck) (not shown) because /u/ is generally considered a back vowel. The notation \$sil<th means the left part of the phone th is in the context of “silence” (\$sil); th>\$fnt means the right part of the phone th is in the context of a front vowel.

3.2. Baseline Training Method

As a first step in training the neural network, Perceptual Linear Prediction (PLP) [5] features (including energy) are computed at non-overlapping 10-msec frames. At each frame of interest, a vector of PLP features is constructed using surrounding frames; we use PLP features from frames at -80,

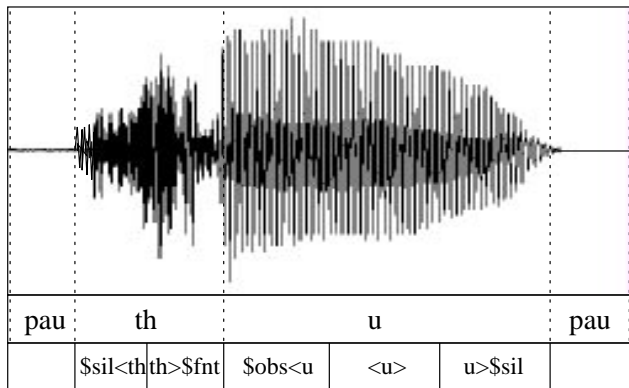


Figure 1. Segmentation of the word “two” using the baseline method.

-40, -10, 0, 10, 40, and 80 msec relative to the frame of interest.

The training data are searched to find 2000 vectors of each category, if possible. If less than 50 vectors can be found, then additional vectors are taken from training data in the OGI Names corpus [2]. The one exception is silence, which is trained using 8000 vectors, all from the OGI Numbers corpus.

The neural network is trained using the back-propagation method with 56 inputs (8 PLP features at seven frames), 224 nodes in the single hidden layer, and one node for each context-dependent category in the output layer. Training is done for 150 iterations, and the iteration with the best performance on the development test set is chosen to be the final neural network.

3.3. Baseline Recognition Method

For recognition of an utterance, PLP vectors are computed in the same way as for training. These PLP vectors are input to the neural network, which computes for each frame the probabilities that the current frame contains each of the specified categories. The result of classification is therefore a $C \times F$ matrix of probabilities, where C is the number of categories and F is the total number of frames. This matrix is then used by a Viterbi search algorithm to determine the most probable sequence of words.

The Viterbi search uses minimum and maximum durations of each category to constrain the possible word choices, but these are not “hard” limits. If the duration of a hypothesized category falls beyond one of the specified limits, a penalty is applied; this penalty is proportional to the time difference between the specified limit and the hypothesized duration. Initial values for these limits are taken from the hand-labeled speech that is used for training. These values are refined dur-

ing the development stage by evaluating word-level performance on a development test set.

4. THE MODIFIED-DIPHONE RECOGNITION SYSTEM

The modified-diphone system is constructed in almost the same way as the baseline system; the only differences are the way in which the speech is segmented, the categories used, the construction of the PLP vectors, and the word models.

4.1. Diphone Segmentation Method

When segmenting speech for training in the proposed modified-diphone system, the first step is to look at the length of the current phone. If the length of the phone is 120 msec or less, then we split the phone in half. If the length of the phone is greater than 120 msec, then we divide the phone into three parts, with the left division occurring 60 msec after the start of the phone and the right division occurring 60 msec before the end of the phone. This leaves a segment in the middle that is, ideally, not much influenced by context. The length of this middle segment depends on the length of the phone being segmented.

For the purposes of nomenclature, we will say that a phone that has been divided into three parts has a left-most division located 60 msec after the start of the phone; a phone that has been divided into two parts has its left-most division located at the mid-point of the phone. The right-most division of a three-part phone is located 60 msec before the end of the phone, and the right-most division of a two-part phone is located at the mid-point of the phone. Diphones are then constructed by taking the segment of speech from the right-most division of the preceding phone to the left-most division of the current phone, and by taking the segment from the right-most division of the current phone to the left-most division of the following phone.

A segmentation of the same word “two” using this modified-diphone method is illustrated in Figure 2. Note that broad categories are no longer used, and we represent a diphone by writing $phone_1 > phone_2$. The context-independent middle section is represented by $<phone>$.

4.2. Construction of the PLP Vector

In the baseline system, frames at -80, -40, -10, 0, 10, 40, and 80 msec relative to the center frame are taken to construct the PLP feature vector. This gives a vector that has its data clustered around the frame of interest but also includes some outlying frames. In the modified-diphone system, we are interested in how the PLP values change over time, and so we use frames that are evenly spaced; in our experiments we used frames at -60, -40, -20, 0, 20, 40, and 60 msec relative to the center frame.

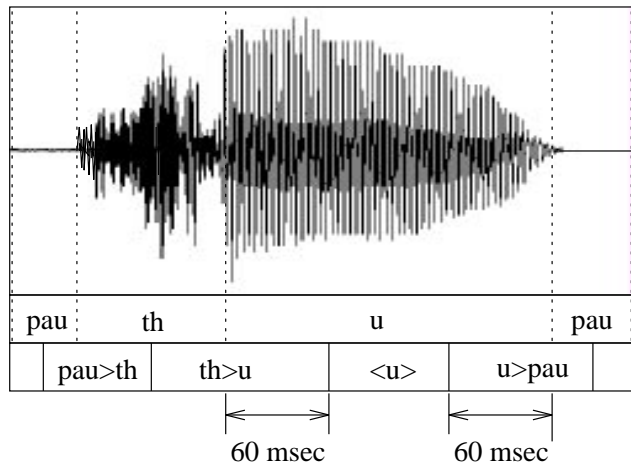


Figure 2. Segmentation of the word “two” using the modified-diphone method.

system	accuracy	subst.	ins.	del.
baseline	94.71%	2.36%	1.88%	1.04%
diphone	95.39%	2.07%	1.34%	1.20%

Table 1. Results for both systems (accuracy, substitutions, insertions, and deletions).

4.3. Diphone Word Models

The word models in the proposed system are constructed so that phones that are usually long in duration (about three-quarters of all phones in the digits task) have an optional context-independent category for the middle part of the phone. Short phones do not have such a middle category. Some short words, such as “two”, require the middle category of a long phone. This is done to prevent insertion errors, although the minimum duration of this middle category can be fairly short.

5. RESULTS

The baseline system required 153 context-dependent categories, and a total of 179,603 vectors were used in training. The diphone system required 150 diphone categories and was trained with a total of 150,684 vectors. Frame-level performance was 56.0% for the baseline system and 51.6% for the diphone system.

For the 1899 digit-string utterances in the test set, the baseline system had word-level performance of 94.71%, and the proposed diphone system had word-level performance of 95.39%; this corresponds to a reduction in error of 12.9%. Table 1 also shows the percentage of substitutions, insertions, and deletions for both systems.

In evaluating the effectiveness of the optional

middle part, we found a 40% reduction in error when using the optional middle part as opposed to using a network trained only on standard diphones.

6. CONCLUSION AND FUTURE WORK

The proposed system yields about a 13% reduction in error compared to the baseline system. These results are encouraging, but the number of diphone categories will increase dramatically for vocabulary-independent systems. A means of reducing the number of categories for such systems must be found. The authors would also like to apply the training technique developed by Yan et al. [6] to the diphone system.

In conclusion, we feel that the modified-diphone technique described here represents an advancement over current biphone and triphone methods for the continuous-digits task.

7. ACKNOWLEDGMENTS

This research was supported in part by grants from the NSF, ONR, DARPA, and member companies of the Center for Spoken Language Understanding.

REFERENCES

- [1] E. Barnard, R. A. Cole, M. Fanty, and P. Vermeulen. Real-world speech recognition with neural networks. *Applications and Science of Artificial Neural Networks*, pages 524–537, April 1995.
- [2] R. A. Cole, M. Noel, T. Lander, and T. Durham. New telephone speech corpora at csu. *Proceedings of the Fourth European Conference on Speech Communication and Technology*, pages 821–824, September 1995.
- [3] W. M. Fisher and J. G. Fiscus. Better alignment procedures for speech recognition evaluation. *International Conference on Acoustics, Speech, and Signal Processing*, 2:II59–II62, 1993.
- [4] J. Schalkwyk, D. Colton, and M. Fanty. The **cslush** toolkit for automatic speech recognition. *OGI Technical Report No. CSLU-011-95*, December 1995.
- [5] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.
- [6] Y. Yan, M. Fanty, and R. A. Cole. Speech recognition using neural networks with forward-backward probability generated targets. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1997. Accepted for publication.