PHONETIC CLASSIFICATION ON WIDE-BAND AND TELEPHONE QUALITY SPEECH

Benjamin Chigier

Speech Technology Group
Artificial Intelligence Laboratory
NYNEX Science and Technology
White Plains, NY, 10604, U.S.A.

1. ABSTRACT

Benchmarking the performance for telephone-network-based speech recognition systems is hampered by two factors: lack of standardized databases for telephone network speech, and insufficient understanding of the impact of the telephone network on recognition systems. The N-TIMIT database was used in the experiments described in this paper in order to "calibrate" the effect of the telephone network on phonetic classification algorithms. Phonetic classification algorithms have been developed for wide-band and telephone quality speech, and were tested on subsets of the TIMIT and N-TIMIT databases. The classifier described in this paper provides accuracy of 75% on wide-band TIMIT data and 66.5% on telephone quality N-TIMIT data. Overall the telephone network seems to increase the error rate by a factor of 1.3.

2. INTRODUCTION

Researchers typically make use of standardized databases in order to benchmark the performance of speech recognition/understanding systems between and within laboratories. Comparisons between laboratories are important in order to benchmark the progress of the field in general. Comparisons within a laboratory are important to benchmark progress as a function of the research-and-development cycle. Benchmarking phonetic classification algorithms for telephone-network-based speech recognition/understanding systems poses two problems. First, there is no commonly-accepted standard database for evaluating phonetic classification for telephone quality speech. As a result, few if any inter-laboratory comparisons have been made. Second, the telephone network presents speech recognition/understanding systems with a band-limited, noisy, and in some cases distorted speech signal. While we would like to benchmark the performance of recognition systems intended for network speech against that of systems intended for wide-band speech, we do not have adequate quantification of the impact of the telephone network's signal degradation on the performance of phonetic classification algorithms. Therefore, we do not know whether the performance of a telephone-speech classification algorithm is limited by characteristics of the algorithm(s) or by characteristics of the test utterances themselves.

Both problems noted above could be addressed given a standardized database in which the speech data is presented in two forms: speech with wide-band characteristics and the same speech data with telephone network characteristics. As reported in Jankowski *et al.* [1], the N-TIMIT database was created for this purpose. The N-TIMIT database is identical to TIMIT [2] except that the former has been transmitted over the telephone network. Figure 1 shows a sample spectrogram of a TIMIT and N-TIMIT utterance. The N-TIMIT versions were recorded over many different transmission paths in order to get a representative sample of the range of telephone network conditions. This data presents a platform to "calibrate" the impact of the telephone network on the performance of phonetic classification algorithms.

The telephone network affects the speech signal it carries in many ways. Chigier and Spitz[3] discussed the possible effects of source characteristics (how the speech is produced) and transmission characteristics (the environment in which the speech is produced, including ambient noise levels and the characteristics of the channel through which the speech is recorded). Some of the more obvious changes are due to band limitation (the signal is band-passed between approximately 300 Hz and 3400 Hz), addition of noise (both switching and line noise), and crosstalk. The goal of this experiment is to quantify the combined effects of signal changes due to telephone transmission characteristics on phonetic classification and to present the performance of a classifier under development. In doing this we hope to provide a model for inter-laboratory and intra-laboratory benchmarking of telephone-based vs. wide-band algorithms.

3. EXPERIMENTAL DESIGN

The wide-band speech data used in this experiment consists of a subset of utterances from the TIMIT database[2]. In

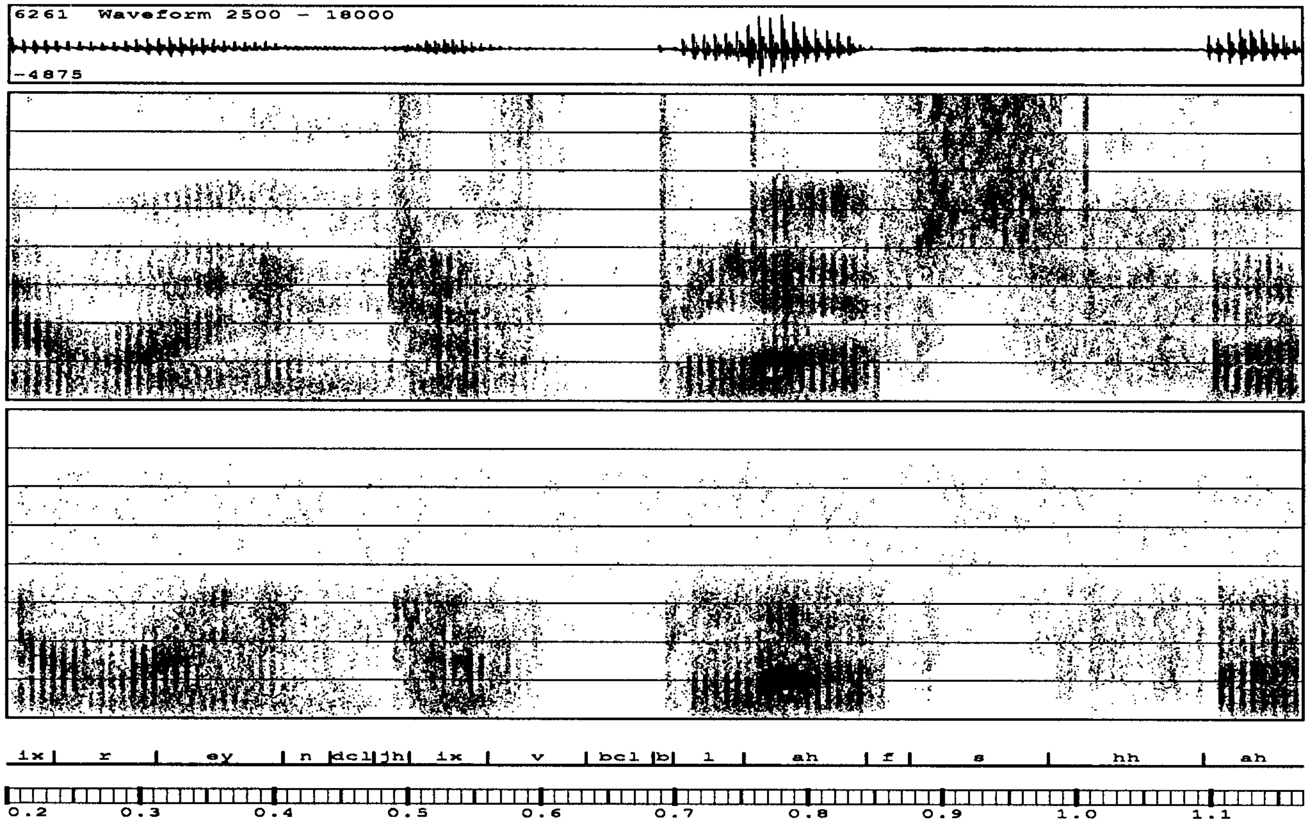

Figure 1: Spectrograms of an utterance in TIMIT (above) and N-TIMIT (below).

order to investigate the effects of telephone network transmission characteristics, the same subset of the TIMIT utterances used by Lee[5], and their N-TIMIT counterparts, were selected. Specifically, the test set consisted of three sentences selected at random from the Brown Corpus[4] ("si" utterances), and five sentences that provide a wide coverage of phoneme pairs ("sx" utterances), all for each of 20 speakers. This resulted in 6016 phonetic segments in 160 unique sentences to be classified into one of 39 categories, also defined by Lee. The "si" and "sx" utterances for the remaining 610 speakers were used to train the classification system.

4. SIGNAL PROCESSING

Identical signal processing was performed on TIMIT and N-TIMIT. The speech signals were sampled at 16 kHz and pre-emphasized. We have developed a new signal representation: bark auditory spectral coefficients (BASC). The BASC was obtained by filtering the FFT representation with the filters of Seneff's auditory model[6]. Specifically, a 128-point FFT was performed with a 28-ms Hanning window every 5 ms. The window size of 28 ms was empirically determined to be the best for this task given this classification system. Each spectral slice, produced by the FFT, was down-sampled to 40 coefficients by computing the dot product between the spectral slice and each of the 40 frequency responses of Seneff's auditory model[6]. This is similar to passing the signal through a bank of critical-band filters.

5. CLASSIFICATION

A full-covariance gaussian classifier was then used to classify each of the incoming segments into one of the 39 phonemes. The gaussian classifier used 56 context-independent models based on a uni-gram model for the phonemes.

5.1. Feature Extraction

Each segment was divided in time into three equal parts. Forty coefficients were averaged across each third, resulting in 120 features for each phoneme. The average spectral difference was computed with its center at the begin boundary and then calculated again with its center at the end boundary. This spectral difference measure was computed for each spectral coefficient (there are 40 spectral coefficients) around each boundary in a segment. Therefore this gave a total of 80 spectral difference features. In calculating the spectral average, the frames further away from the center of the boundary were weighted more heavily than the frames close to the boundary. This weighting scheme is similar to that proposed by Rabiner *et al.* [7]. Let $S[f,c]$ be the value of the spectral representation at frame f and spectral coefficient c. Thus, the spectral difference coefficient at a segment boundary, sb (begin or end boundary), $\Delta S[sb,c]$ is defined as:

$$\text{Eq. 1: } \Delta S[sb,c] = \frac{1}{N}\left\{ \sum_{w=1}^{N} w\left(S[(sb-w),c] - S[(sb+w),c]\right) \right\}$$

where $2N$ is the number of frames in the overall window, and w is the weighting factor.

A pilot study was conducted to determine whether weighted averages provide better classification performance than traditional unweighted (the special case of $w = 1$ in Eq. 1) averages using the current classification system. The weighted versions slightly outperformed the unweighted averages when testing on the cross-validation set described above.

Another pilot study was designed to determine the optimal number of frames to use when computing the weighted averages. The number of frames included was systematically varied from 0 to 10 ($0 \leq N \leq 10$ in Eq 1), both preceding and following the boundary, which resulted in a weighted average difference for each coefficient. (Note that for $N = 0$ frames, no difference information is derived). The optimal number of frames to include in the weighted average was found to be 7, which provided the highest classification score on the cross-validation set.

The average spectral distance calculations result in 40 features at the begin boundary and 40 features at the end boundary. These were combined with the 120 features derived for each segment described above. Duration and maximum zero crossing count were added to the pool of features, resulting in 202 features that were passed on to the classification system.

5.2. Feature Selection

Principal component analysis was used to reduce the number of input dimensions to the classifiers. The principal components were ranked in decreasing order according to amount of variance accounted for in the original data (i.e., based on the eigenvalues). The final set of principal components used was determined empirically by adding one principal component at a time to the classifier, training the classifier, and then evaluating performance on the cross-validation set. Finally, the set of principal components that produced the best performance on the cross-validation set was used to train the classifier on the entire training set. This procedure was carried out separately for the N-TIMIT and the TIMIT database. The resulting two classifiers were evaluated on their respective test sets.

Ranking the principal components according to the amount of variance they account for may not reflect how well they discriminate between classes. Therefore, another procedure was also evaluated to determine which of the principal components have the most discriminating power. This procedure was a stepwise add-on procedure based on adding the principal component that improves the performance of the classifier the most on the cross-validation set. This ranking of the principal components was determined by first training a classifier on all 202 principal components. Another classifier was then created by taking the features from the initial classifier, one at a time, and testing on the cross-validation set. The principal component that performed the best was next used with the remaining features one at a time, and now the pair of features that gave the best performance was used with the remaining features. This procedure was carried out by incrementally adding principal components to the classifier based on their ability to improve performance. This procedure is not an optimal procedure, but it is computationally feasible (the optimal procedure would require testing 2^{202} (approximately $6.4\text{x}10^{60}$) classifiers).

6. RESULTS

The eigenvectors are ordered according to the amount of variance that they account for in the original feature space; we can therefore draw a plot of the percentage of the total variance the principal components account for of the original data as the number of principal components increases. Figure 2 displays the number of principal components in the system vs. the percentage of the total variance that is accounted for by those principal components. In N-TIMIT,

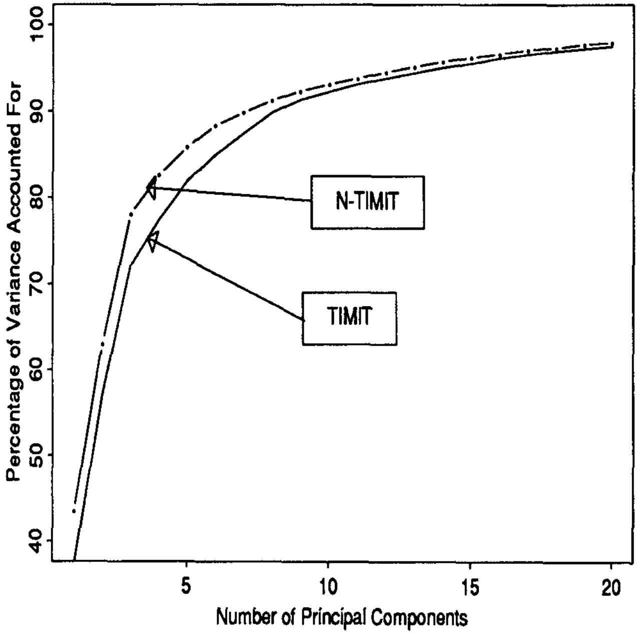

Figure 2: Number of principal components used vs. the percentage of variance accounted for by the principal components.

information in the spectrum above 3400 Hz is small (due to the bandpass characteristics of the telephone network) and so the variance of the features that represent this information is small. Consequently fewer principal components are needed to account for the variability of these features. This can be seen in Figure 2, where the N-TIMIT curve is higher than the TIMIT curve. A larger percentage of the variance is accounted for in N-TIMIT than in TIMIT for the same number of eigenvectors.

Figure 3 is a plot of TIMIT error rate and N-TIMIT error rate on the cross-validation set. It is interesting to note that after the top 10 principal components have been used, the mean value of the ratio of N-TIMIT error rate to TIMIT error rate is 1.3, with a standard deviation of only 0.019. The error rate with 10 principal components is 39.6% and 48.1% for TIMIT and N-TIMIT respectively and goes down to a minimum of 25.8% and 34.1% for TIMIT and N-TIMIT respectively on the cross-validation set. The number of principal components discovered to give the best classification performance on the cross-validation set was 58 for the TIMIT classifier and 65 for the N-TIMIT classifier. The improvements in classification accuracy, however are very small after approximately 35 principal components have been included.

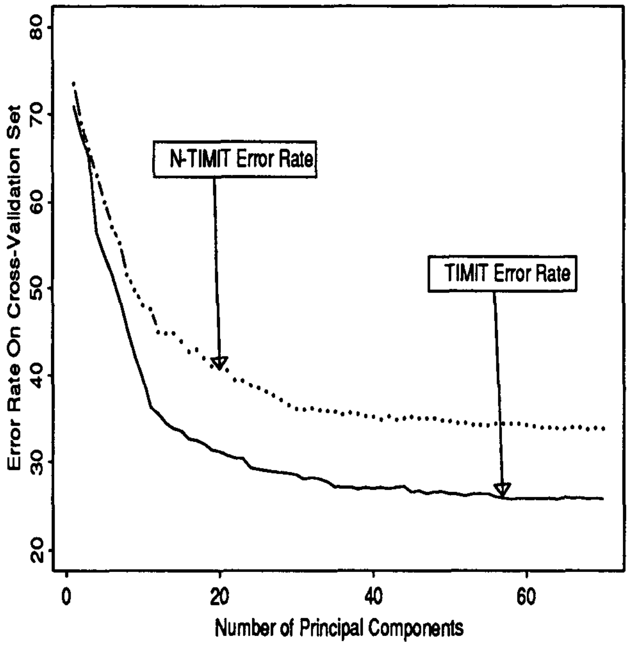

Figure 3: Number of principal components used vs. error rates for TIMIT and N-TIMIT classifiers on the cross-validation set.

The two procedures for ranking the principal components were compared. The first procedure ranked the principal components according to the variance they accounted for; the second ranked them according to their discriminative power. No difference in classification accuracy was found between these two procedures. This finding concurs with Brown[8]; The performance of his system when a large number of principal components was used was the same as when he used discriminative analysis.

The first-choice accuracies of the TIMIT and N-TIMIT classifiers on the test set are 74.8% and 66.5% respectively. Error rates of the two classifiers on the test set appear in Table 1. As on the cross-validation set, the phonetic classification error rate on the test set is also increased by a factor of 1.3 by the telephone network. In order to determine whether TIMIT and N-TIMIT classification accuracy differ significantly, a McNemar test of symmetry was conducted. The results of this analysis revealed significant differences between TIMIT and N-TIMIT classifier performance ($p < 0.01$).

Database	First Choice Error Rate	Top 2 Choices Error Rate	Top 3 Choice Error Rate
TIMIT	25.2%	11.6%	6.4%
N-TIMIT	33.5%	17.5%	11.4%

Table 1: Error rates on test set.

A McNemar test of symmetry was also conducted separately on each of the 39 phonemes to determine which phonemes accounted for the significant differences. The results of this analysis revealed a significant effect of database on 13 of the 39 phonemes ($p < 0.01$). These phonemes are shown in Table 2. The percentage of N-TIMIT phonemes correctly classified were subtracted from the percentage of TIMIT phonemes correctly classified. Results are presented in decreasing order. For example, the accuracy on the pho-

Phoneme	Difference in % correct	Difference in # correct
f	29	25
g	28	14
k	27	45
s	22	55
hh	20	13
m	18	24
r	17	34
p	17	19
z	15	22
l	13	29
er	12	23
n	8	30
cl	3	38

Table 2: Phonemes that are significantly different ($p < 0.01$) between TIMIT and N-TIMIT.

neme, /f/, is 29% higher on TIMIT than on N-TIMIT. A large number of these errors are predictable based on the acoustic characteristics of the segments and their sensitivity to band-passing or noise. A spectrogram of the same TIMIT and N-TIMIT utterance is shown in Figure 1. This utterance was chosen because it highlights several of the phonemes that are classified significantly differently in TIMIT and N-TIMIT. Many of the classification errors are explainable from the spectrogram. The frication for /s/, for example, is a visible and salient cue in the TIMIT utterance, but is nearly non-existent in the telephone quality N-TIMIT version.

7. CONCLUSIONS

In developing the TIMIT and N-TIMIT classifiers described in this experiment, three findings emerged for improving classification performance:

1. A Hanning window size of 28 ms was determined to be the best for this task;
2. Weighted average spectral differences outperformed unweighted averages;
3. The number of frames to include in the weighted average spectral difference was found to be 7.

The advantage of weighted spectral differences is supported by earlier results reported by Rabiner *et al.* [7]. It remains to be seen whether the characteristics determined in this setting will transfer well to other recognition tasks.

The performance of the TIMIT classifier (75%) compares favorably to results reported by other researchers[5,9,10,11]. Results indicate that the telephone network, in general, increases the phonetic classification error rate by a factor of 1.3. This correction factor may be useful in our attempts to benchmark the performance of wide-band vs. network based recognition systems. Furthermore, this study sets a first benchmark on the N-TIMIT database. We hope to encourage others to evaluate their systems on this database and in so doing follow the model established by our colleagues working on wide-band speech.

REFERENCES

1. Jankowski, C., Kalyanswamy, A., Basson, S., and Spitz, J., "N-TIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database", *ICASSP,* 1990, pp. 109-112.

2. Fisher, W., Doddington, G., and Goudie-Marshall, K., "The DARPA speech recognition database: specifications and status", *DARPA Workshop on Speech Recognition,* Feb 1986, pp. 93-99.

3. Chigier, B., and Spitz, J., "Are laboratory databases appropriate for training and testing telephone bandwidth speech recognizers?", *ICSLP,* 1990, pp. 1017-1020.

4. Kucera, H., and Francis, W. N., "Computational analysis of present day American English", *Brown University Press,* Providence, RI, 1967.

5. Lee, K,. Hon, H., "Speaker independent phone recognition using hidden Markov models". *Carnegie-Mellon University, Computer Science Dept,* Ref Number CMU-CS-88-121 March 1988.

6. Seneff, S., "A Joint Synchrony/Mean Rate Model of Auditory Speech Processing", *Journal of Phonetics,* Vol. 16, No.1, pp 55-76, 1988

7. Rabiner, L. R., Wilpon, J. G., and Soong, F. K., "High performance connected digit recognition, using hidden Markov models.", *ICASSP,* 1988, pp. 119-122.

8. Brown, P. F., "The acoustic-modeling problem in automatic speech recognition", Ph. D. Thesis, Carnegie-Mellon University, March 1988.

9. Zue, V, Glass, J., Phillips, M., and Seneff, S., "Acoustic Segmentation and Phonetic Classification in the SUMMIT System", *ICASSP,* 1989, pp. 389-392.

10. Leung, H. C., Glass, J. R., Phillips, M. S., and Zue, V. W., "Detection and Classification of Phonemes Using Context-Independent Error Back-Propagation", *ICSLP,* 1990, pp. 1061-1064.

11. Digalakis, V., Rohlicek, J. R., and Ostendorf, M., "A dynamical system approach to continuous speech recognition", *DARPA, Speech and Natural Language Workshop, Feb 1991, pp 253*-257.