Analysis of Confusion Matrix to Combine Evidence for Phoneme Recognition

S. R. Mahadeva Prasanna [a]
B. Yegnanarayana [b]
Joel Praveen Pinto and Hynek Hermansky [c] [d]

IDIAP–RR 07-27

July 2007

SUBMITTED FOR PUBLICATION

[a] Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, Guwahati-781039, Assam, India, email: prasanna@iitg.ernet.in
[b] International Institute of Information Technology Hyderbad, India, email: yegna@iiit.ac.in
[c] IDIAP Research Institute, Martigny, Switzerland
[d] Swiss Federal Institute of Technology (EPFL), Switzerland, email:{jpinto,hynek}@idiap.ch

Analysis of Confusion Matrix to Combine Evidence for Phoneme Recognition

S. R. Mahadeva Prasanna B. Yegnanarayana
Joel Praveen Pinto and Hynek Hermansky

July 2007

submitted for publication

Abstract. In this work we analyze and combine evidences from different classifiers for phoneme recognition using information from the confusion matrices. Speech signals are processed to extract the Perceptual Linear Prediction (PLP) and Multi-RASTA (MRASTA) features. Neural network classifiers with different architectures are built using these features. The classifiers are analyzed using their confusion matrices. The motivation behind this analysis is to come up with some objective measures which indicate the complementary nature of information in each of the classifiers. These measures are useful for combining a subset of classifiers. The classifiers can be combined using different combination schemes like product, sum, minimum and maximum rules. The significance of the objective measures is demonstrated in terms the results of combination. Classifiers selected through the proposed objective measures seem to provide the best performance.

1 Introduction

In Automatic Speech Recognition (ASR) an important objective is to obtain information about the underlying phoneme sequences from the acoustic signal. For this, the acoustic signal is first transformed into a sequence of feature vectors using different representations of speech signal like Mel-Frequency Cepstral Coefficients (MFCCs), Perceptual Linear Prediction (PLP) and Multi-RASTA (MRASTA) schemes [1–3, 13]. The extracted features are then used for building classifiers, and these classifiers are used for recognizing the phonemes. The performance of the classifiers depends both on the type of feature selected, and also on the approach taken for building the classifier. However, in practice, it is observed that any one type of feature or approach for building a classifier may not yield the best performance. Attempts have been made to build multiple classifiers for the same acoustic signal using different features and classifiers and then combine them to obtain a better classifier. It is known that careful choice of features and individual classifiers may yield better performance than the individual classifier [4–6]. To choose the individual classifiers, one thumb rule is to group only those classifiers which have complementary information. How to know the complementary nature of information in the classifier? One immediate answer is to select features obtained from different signal processing methods and/or classifiers [7].

Alternatively, if we are given only a set of classifiers, is it possible to derive some objective measures to indicate the complementary nature of information among the selected subset of classifiers? This may help us in selecting a subset of classifiers without having to know the details of the classifier. We believe that the confusion matrix may provide valuable information for detecting the complementary nature of information present in a set of classifiers. The objective measures derived helps us to take decision to include or not an individual classifier for obtaining the combined classifier. The rest of the paper is organized as follows: Section 2 describes the different classifiers used in the present work. These include one frame PLP, nine frame PLP, MRASTA and context based phoneme classifiers [7]. Section 3 describes the analysis of the confusion matrices from different classifiers from the point of phoneme recognition performance. Objective measures are proposed for measuring the complementary information. Section 4 describes the selection of a subset of classifiers using the proposed objective measures. The effectiveness of the proposed objective measures is demonstrated in terms of the performance of the combined classifier. The work is summarized in Section 5, and scope for future work is discussed in this section.

2 Classifiers for Phoneme Recognition

2.1 Database

We have used TIMIT speech database for this study [8]. It consists of speech from 630 speakers, out of which 438 are male and 192 are female speakers. Each of the speakers have spoken 10 speech sentences, out of which 2 sentences are common across all the speakers. Therefore, in total we have 6300 sentences, and out of this about 70% of the data is used for training, and rest is used for testing. To avoid biasing during training and testing, the common sentences are removed from the database, and hence we are left with 5040 speech utterances, out of which 3696 speech utterances are used for training, and 1344 are used for testing. The TIMIT database is hand labeled using 61 phonetic labels, and are merged into 39 labels in the present work as discussed in [9,11]. The label representing silence region /sil/ is not used in evaluating the classifier, since it occurs maximum number of times, and hence is well trained. Thus we are left with 38 classes for the study. We are building 39 class classifier, but using output only from 38 classes for further analysis and combining.

2.2 One Frame PLP (PLP-1) based phoneme recognition system

Speech signal is processed by the PLP analysis in frames of 25 msec with a shift of 10 msec to extract 13 PLP based cepstral coefficients, and their first and second order derivatives for each frame. Thus we have feature vectors of 39 dimension for building the phoneme classifier. A Multi-Layer Perceptron (MLP) of three layers with dimensions 39:1000:39 is trained using the PLP feature vectors extracted from the 3696 speech utterances [12]. The trained MLP is tested using 39 dimension feature vectors extracted from the 1344 test speech utterances. The trained MLP gives posterior probabilities for each of the 39 classes, which are then decoded in the hybrid HMM/ANN approach to find the phoneme sequences [11]. The phoneme recognition performance of the system is given in Table 1. The average phoneme recognition accuracy given in the table is computed using the relation

$$R(\%) = (1 - (I + S + D)/N) \times 100 \tag{1}$$

where, R is the average phoneme recognition accuracy, I, S, D and N are the total number of insertions, substitutions, deletions and actual phoneme examples, respectively, across the 38 classes for all the testing speech data.

2.3 Nine Frame PLP (PLP-9) based phoneme recognition system

To improve the performance of the phoneme recognition system, one approach followed is to exploit the context information by training the neural network for each frame using context of a few frames and in the present work a context of 9 frames is used [3, 7]. The motivation for this approach is that, the neural network not only adjusts the weights according to the spectral information available in the current frame, but also according to the spectral information available in the adjacent frames. The network structure used is 351:1000:39. The input dimension 351 is due to nine frames (39 × 9). The phoneme recognition performance of the system is given in Table 1. As it can be observed, the PLP-9 classifier shows a significant improvement in performance compared to PLP-1 classifier.

2.4 MRASTA based phoneme recognition system

Multi-RASTA features also have been proposed as robust features for ASR tasks [3]. Since they are derived over long term temporal trajectories, they exhibit robustness to channel distortions. Recently it is also demonstrated that MRASTA has complementary information compared to the PLP features [7]. Hence classifiers can also be built using MRASTA as features. The speech signals are processed in frames of 25 msec with a shift of 10 msec to extract 19 critical band energies for every 10 msec. The critical band energies are temporally filtered by a set of 16 first and second order Gaussian derivatives. The 16 filters differ in the temporal resolution achieved by varying the width/spread of the Gaussian function [11]. The temporally filtered sequences are then subjected to first order frequency domain filtering. In this way we will have 576 ($16 \times 19 + 17 \times 16$) features for every 10 msec. The MLP of structure 576:1000:39 is trained using the MRASTA features derived from the training data. The phoneme recognition performance of this system is given Table 1. The performance of MRASTA is significantly better than the PLP-1 case, but lower than the PLP-9 case. However, as it will be demonstrated later, since it is built using different feature compared to PLP, it pairs up well with other classifiers to provide improved performance in the combined system.

2.5 Context-based Phoneme recognition system

When we consider PLP-9 based phoneme recognition system, the PLP features are extracted from all the frames of the given phoneme speech, and the MLP is trained. Some phonemes may have longer duration and different acoustic characteristics at different locations, and hence training only one MLP by using all the frames may not able to effectively capture all the information present in each phoneme. Alternatively, each phoneme duration may be divided into three equal regions, one indicating the left context of phoneme, the second indicating the right context and the third representing the middle context. Separate models may be trained for each context, and the evidences may be further combined to obtain phoneme class [10, 11]. Three MLPs with structure 351:1000:39 are trained. The evidence available at the output of each classifier are then merged to get phoneme evidence in a nonlinear fashion using another MLP of dimension 120:3000:39. The phoneme recognition performance of this system is given in Table 1. As it can be observed from the table, this classifier provides slightly improved performance over PLP-9 classifier.

Table 1: Phoneme Recognition Performance of Different Phoneme Classifiers. In total there are 42557 phoneme examples (N) in the testing data for the 38 classes considered. In the table the abbreviations A, I, D and S indicate the number of phonemes detected, inserted, deleted and substituted and R indicate the phoneme recognition accuracy.

Type of Classifier	A	I	D	S	R
PLP-1	27747	1388	5821	8989	61.94%
PLP-9	30111	1500	4686	7760	67.23%
MRASTA	29126	1798	4884	8574	64.22%
CONTEXT	31168	2226	3578	7811	68.01%

3 Analysis of Confusion Matrix

Complementary information may be available in terms of features and/or classifiers. If we have access only to the classifier outputs, then how to select a subset of classifiers for combination? The output of the classifier in the form of confusion matrix provides valuable information to indicate its behavior with respect to the classification task. We may also find some relation between the two classifiers by visually comparing their confusion matrices. However, it will be nice to have some objective measures obtained from the confusion matrix, which can give information about the classifier and also pair of classifiers. Some objective measures may be derived for each classifier, and also for a pair of selected classifiers. In this work, the objective measures which indicate information about the properties of a given confusion matrix are termed as *Intra-Confusion Objective Measures*, and those which indicate about properties of a pair of confusion matrices are termed as *Inter-Confusion Objective Measures*.

3.1 Intra-Confusion Objective Measures

3.1.1 Classification Accuracy

The classification accuracy of a classifier can be measured as the ratio of the sum of principal diagonal values to the total sum of values in the confusion matrix. If C indicate the confusion matrix, the classification accuracy A_c can be defined as

$$A_c = \frac{\sum_{i=1}^{N} C_{ii}}{\sum_{i=1}^{N}\sum_{j=1}^{N} C_{ij}} \tag{2}$$

where N is the number of rows or columns in the confusion matrix. For given task, the classifier should have as high values of A_c as possible. Since this seldom happens in practice, classifiers with higher values of A_c are preferred. This is desirable, if you want to select the best classifier for grouping. Alternatively, if you want to choose more than one classifier for further combination, then other factors also need to be considered.

3.1.2 Individual Class Accuracy Variance

The classification accuracy A_c gives information only about the average performance of the classifier. It is possible that relatively high value of A_c is due to very high accuracy among some classes, and very low among some other classes. However, what is desirable from the classifier is, it should provide high accuracy for all classes or at least the performance for all classes should be around the mean value A_c. This information can be obtained by computing the individual class accuracy variance. If V_c indicates the individual class accuracy variance, then it can be defined as

$$V_c = \sum_{i=1}^{N} (\frac{C_{ii}}{\sum_{j=1}^{N} C_{ij}} - A_c)^2 \tag{3}$$

Thus a classifier with lowest variance for individual class accuracy is desirable in practice.

3.1.3 Weighted Confusion Index

In an ideal condition, there should not be any confusion among different classes. However, as the name of the matrix indicates, there will be always confusion among different classes, and are indicated by non-zero off diagonal entries. The summation of non-zero off diagonal elements weighted by the maximum value (excluding the principal diagonal element) in each row gives an average information about the confusion pattern exhibited by the classifier. The weighted confusion index I_c can be defined as

$$I_c = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{C_{ij}}{max\{C_{ij}\}} \qquad for i \neq j \tag{4}$$

In the ideal condition I_c will be zero. However, in practice, a classifier with I_c value as low as possible is desirable.

3.1.4 Symmetric Class Confusion Difference

As mentioned earlier, there will be confusion among classes. However, for a given classifier the confusion may have some regular structure. For instance, most of the examples of class i may be confusing with j, and vice versa. Even though, such confusions are high, they can easily be merged or train the classifier better by using additional information in the feature space, which distinguishes the two classes. The classifier confusion behavior may therefore be computed by neglecting such high numbers. One easy way is to consider the absolute difference between the i^{th} row and column, and sum them. By doing this, we will be computing the difference among the classes which are symmetric to the principal diagonal, and hence this measure is termed as symmetric class confusion difference (D_c). It may be defined as

$$D_c = \sum_{i=1}^{N} \sum_{j=1}^{N} |C_{ij} - C_{ji}| \tag{5}$$

By the way of computing, the high values which may be present among the symmetric classes will be minimized, and hence a classifier with lower value of D_c is preferable in practice.

3.2 Inter-Confusion Objective Measures

3.2.1 Dissimilarity of Confusion Patterns

From the combination point of view, the two confusion matrices which have relatively less similarity among the confusion patterns are preferred. This is because, if one classifier confuses, other should not, then only they can aid each other to reduce confusion, when they are combined. A measure can be computed for the same by considering respective rows in the two confusion matrices, which is termed as dissimilarity of confusion patterns. The i^{th} row of the confusion matrix C^x is compared with the i^{th} row of the confusion matrix C^y by using the correlation coefficient. Since the maximum value for the correlation coefficient is unity, the similarity value subtracted from unity should give a measure about the dissimilarity of the two patterns. This value accumulated over all the rows will give a measure of dissimilarity of confusion patterns. If D_{xy} indicates the dissimilarity measure of confusion patterns, then it can be defined as

$$D_{xy} = \sum_{i=1}^{N} 1 - \frac{C_v(C_i^x, C_i^y)}{S_d(C_i^x) S_d(C_i^y)} \tag{6}$$

where, C_v is covariance, and S_d is standard deviation. If the two confusion patterns for comparison are exactly similar, then the correlation coefficient will be unity, then D_{xy} will be zero. Accordingly, we desire a higher value for D_{xy}. Hence a classifier pair X and Y having relatively more value for D_{xy} is preferred for combination.

3.2.2 Distance of Confusion Patterns

From the combination point of view, two confusion matrices which have relatively more distance among the confusion patterns are preferred. This is because, if the distance is large, then it indicates that the patterns are not close in the confusion space, and hence when they combine, they may be minimized, provided suitable combination scheme is used. A measure can be computed by considering respective rows in the two confusion matrices, which is termed as distance of confusion patterns. The Euclidian distance between the i^{th} row of the confusion matrix C^x and the i^{th} row of the confusion matrix C^y is computed and accumulated over all the rows which will give a measure of the distance of confusion patterns. If E_{xy} indicates the distance measure of confusion patterns, then it can be defined as

$$E_{xy} = \sum_{i=1}^{N} \sum_{j=1}^{N} (C_{ij}^x - C_{ij}^y)^2 \tag{7}$$

If the two patterns are at the same place in the confusion space, then their Euclidian distance will be zero. Hence, from the combination point the distance between the confusion patterns measured in terms of E_{xy} should be as high as possible.

3.2.3 Binary Difference in Confusion Patterns

From the combination point of view, the difference in confusion pattern between the two confusion matrices should be as high as possible. This is because, it indicates that the two classifiers are different or have different features employed for classification. Hence we may get maximum benefit by combining such classifiers. This can be computed by first converting the given confusion matrices into matrices having only zeros and ones as entries. This is to make sure that the actual values are not important, but only the distribution of their pattern. After this, the difference between the two confusion matrices is taken to find out how different the two matrices confusion pattern. If C^{xb} and C^{yb} are the modified matrices of C^x and C^y having only ones and zeros as entries, and P_{xy} indicates the confusion pattern between the two matrices, then it can be defined as

$$P_{xy} = \sum_{i=1}^{N} \sum_{j=1}^{N} (C_{ij}^{xb} - C_{ij}^{yb}) \tag{8}$$

As per the design of the measure, the two confusion matrices having relatively maximum P_{xy} are preferred for combining.

4 Selection of Classifiers and Combining Evidences

The intra-confusion objective measures defined in the earlier section are computed for the four classifiers described in Section 2 and are listed in Table 2. The inter-confusion objective measures defined in the earlier section computed for the different classifier pairs are listed in Table 3.

Table 2: The intra-confusion objective measures computed for the confusion matrices of different phoneme classifiers. The abbreviations A_c, V_c, I_c and D_c indicate classification accuracy, individual class accuracy variance, weighted confusion index and symmetric class confusion difference, respectively.

Type of Classifier	A_c	V_c	I_c	D_c
PLP-1	63.14%	1.54	114.00	7870
PLP-9	68.52%	0.82	118.32	5132
MRASTA	66.28%	0.91	130.03	5558
CONTEXT	69.60%	0.70	118.65	5430

Table 3: The inter-confusion objective measures computed for the confusion matrices of different pair of classifiers. The abbreviations D_{xy}, E_{xy} and P_{xy} indicate dissimilarity of confusion patterns, distance of confusion patterns and binary difference in confusion patterns, respectively.

Type of Classifier	D_{xy}	E_{xy}	P_{xy}
PLP-1 & PLP-9	2.09	8.71	280
PLP-1 & MRASTA	2.47	10.13	327
PLP-1 & CONTEXT	2.89	10.19	309
PLP-9 & MRASTA	2.56	11.59	309
PLP-9 & CONTEXT	1.80	7.85	265
MRASTA & CONTEXT	3.33	13.25	336

From the point of view of selection of individual best classifier, the intra-confusion measures A_c, V_c, I_c and D_c are preferred. The classifier which gives highest rank for majority of these measures may be chosen as the best classifier. Accordingly, if we observe Table 2, the first best classifier is the *CONTEXT* classifier, the second best is *PLP-9*, the third best is *MRASTA* and the last one is *PLP-1*. Alternatively, when we go for combining the classifiers, then apart from the above measures, we also have to obtain their values for inter-confusion measures D_{xy}, E_{xy} and P_{xy} as given in Table 3. The classifier pair which gives highest rank for majority of these measures may be chosen as the best classifier pair for further combination. As it can be observed from the Table 3, the best classifier team will be formed using *MRASTA* and *CONTEXT*, which is guaranteed to provide the best performance compared to all other classifier teams.

We have combined different classifier teams using simple combination schemes like product, sum, minimum and maximum schemes [4–6]. The combined classifier system results are given in Table 4. As it can be observed from the results, the classifier formed using *MRASTA* and *CONTEXT* yields the best performance, immaterial of whatever the combination used for constructing the combined classifier.

Table 4: The performance of different classifier teams under different combination schemes. The entries in the table are expressed in percentage (%). The abbreviations P, S, M_i and M_x indicate product, sum, minimum and maximum combination schemes, respectively.

Type of Classifier	P	S	M_i	M_x
PLP-1 & PLP-9	66.22	66.54	66.86	65.96
PLP-1 & MRASTA	66.82	66.44	67.31	65.52
PLP-1 & CONTEXT	66.59	67.54	67.80	67.08
PLP-9 & MRASTA	68.25	68.47	68.95	67.76
PLP-9 & CONTEXT	67.34	68.83	68.77	68.44
MRASTA & CONTEXT	68.75	69.25	69.54	68.77

5 Summary and Conclusions

The objective of this work was to analyze the confusion matrices of different classifiers and come up with some objective measures which help in selecting a best subset of classifiers for combination. Accordingly a set of intra-confusion and inter-confusion objective measures are proposed. Using the proposed objective measures a subset of classifiers namely MRASTA and CONTEXT from the whole set is chosen. It is experimentally demonstrated that this pair provides the best performance, immaterial of the combination scheme used.

As it can be observed from the experimental results, the improvement due to combination is not very significant. This is because our objective was only to analyze and develop some objective measures and we have made no attempt to optimize individual classifiers. A careful effort in this direction may further improve the performance of the combined system significantly.

Acknowledgements

This research work is part of Indo-Swiss Joint Research Project (ISJRP) on ”Keyword Spotting in Continuous Speech”. The authors would like to thank Joel Pinto and Fabio Valente all from IDIAP, Martigny, Switzerland. This work was supported from Joint Indo-Swiss Project No.IT03, of DARPA GALE and of EC Integrated Project DIRAC.

References

[1] S. B. Davis and P. Mermelstein : Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust., Speech and Signal Processing, vol. 28, pp. 357-366, 1980.

[2] H. Hermansky : Perceptual linear predictive analysis of speech. J. Acoust. Soc. Amer., vol. 87(4), pp. 1738-1752, 1990.

[3] H. Hermansky and P. Fousek: Multi-resolution RASTA filtering for TANDEM-based ASR, in Proc. INTERSPEECH-2005, (Lisbon, Portugal), 2005.

[4] L. Xu, A. Krzyzak and C. Y. Suen: Methods of combining multiple classifiers and their application to handwriting recognition, IEEE Trans. System, Man, Cybernatics, vol. 22(3), pp. 418-435, 1992.

[5] T. K. Ho, J. J. Hull and S. N. Srihari: Decision combination in multiple classifier systems, IEEE Trans. Pattern Analysis, Machine Intelligence, vol. 16(1), pp. 66-75, 1994.

[6] J. Kittler, M. Hatef, R. P. W. Duin and J. Matas: On combining classifiers, IEEE Trans. Pattern Analysis, Machine Intelligence, vol. 20(3), pp. 226-239, 1998.

[7] S. R. M. Prasanna and H. Hermansky, Multi-RASTA and PLP in Automatic Speech Recognition, Proc. INTERSPEECH-2007, (Antwerp, Belgium), 2007.

[8] W. M. Fisher et al.: The DARPA speech recognition research database: specifications and status, Proc. DARPA Workshop on Speech Recognition, pp. 93-99, 1986.

[9] K-F Lee and H-W Hon: Speaker independent phoneme recognition using hidden Markov models, vol. 37(11), pp. 1641-1648, 1989.

[10] P. Schwartz, P. Matjka and J. Ernock: Hierarichical structures of neural networks for phoneme recognition, Proc. ICASSP, pp. 325-328, 2006.

[11] J. Pinto et. al.: Phoneme recognition on TIMIT using HMM-ANN approach, IDIAP Research Report, no. IDIAP-RR-07-27, 2007.

[12] Haykin, S.: Neural Networks: A Comprehensive Foundation. Macmillan College Publish-ing Company, New York (2004)

[13] Rabiner, L. R., Juang, B.H.: Fundamentals of Speech Recognition. Prentice Hall, Engle-wood Cliffs, NJ, (1993)