

On Estimation of A Speaker’s Confusion Matrix from Sparse Data

Stephen Cox

School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, U.K.

S.J.Cox@uea.ac.uk

Abstract

Confusion matrices have been widely used to increase the accuracy of speech recognisers, but usually a mean confusion matrix, averaged over many speakers, is used. However, analysis shows that confusion matrices for individual speakers vary considerably, and so there is benefit in obtaining estimates of confusion matrices for individual speakers. Unfortunately, there is rarely enough data to make reliable estimates. We present a technique for estimating the elements of a speaker’s confusion matrix given only sparse data from the speaker. It utilizes non-negative matrix factorisation to find structure within confusion matrices, and this structure is exploited to make improved estimates. Results show that under certain conditions, this technique can give estimates that are as good as those obtained with twice the number of utterances available from the speaker.

1. Introduction

The use of confusion matrices in speech recognition has been a recurrent theme, especially in cases where phone (rather than word) recognition is used [1, 2, 3]. Little attention has been paid to the variation in confusion matrices across speakers: most researchers have assumed that a “speaker-independent” confusion matrix averaged over many speakers is a good enough representation of the pattern of errors made by any speaker. The argument might run something like this: the accuracy of a human “recogniser” on the speech of a native speaker of his or her own language approaches 100% at even moderate SNRs [4], so the much lower accuracy of machine recognition is almost entirely attributable to the machine rather than the speaker, and is therefore fixed and independent of speaker. However, an analysis (as part of the experiments reported here) of confusion matrices estimated from utterances provided by 91 speakers revealed large variations from speaker to speaker in the patterns of confusions of phonemes, especially vowels, and this gives us the motivation to model the error-patterns from individual speakers. A speaker-independent confusion matrix can be obtained quite easily from speech from many speakers, perhaps collected over a considerable time. However, collection of enough data from a single speaker to make reliable estimates of his or her confusion matrix is usually difficult, as in practice, only a small amount of data will be obtainable—the same problem is encountered in speaker adaptation. Hence we focus here on the problem of obtaining reliable estimates of a speaker’s confusion matrix using a small amount of data presented to the speech recogniser.

The structure of the paper is as follows: in section 2, we describe the motivation for the approach used here, give a basic account of non-negative matrix factorisation (NNMF) and also present some evidence for structure within confusion matrices. Section 3 describes the data used, and gives details on the composition and estimation of the confusion matrices. Section 4

gives details of the models used and how the experiments were performed, and section 5 presents and discusses results on both training- and test-sets. We end with a discussion and ideas for future work.

2. Background and Approach

We are concerned here with estimating the values of a set of random variables (confusion matrix elements) from a sparse set of examples. A standard way of approaching this problem is to exploit the correlations between the variables in such a way that the estimates are constrained to exhibit a similar set of correlations. The main benefit of this technique is that it is able to remove some of the noise present in the sparse examples. Such techniques have been applied frequently in speech and language processing e.g. for speaker adaptation from sparse data [5], for enhancement of speech [6], for incorporation of semantics into language models [7], for prediction of motion in audio-visual speech [8] etc. In this paper, we utilise the correlations between elements in confusion matrices to estimate an individual speaker’s confusion matrix given some sparse data from that speaker.

There are two main approaches that have been used for these kind of problems, principal component analysis (PCA) and singular value decomposition (SVD), which are closely related to each other. Estimates are made by projecting the sparse data into a subspace defined by the eigenvectors (PCA) or singular vectors (SVD) that have been estimated from many examples in the training data, and then projecting back into the original space. The estimates produced by both these techniques are unbounded. However, the elements of a confusion matrix are probabilities and hence are in the range $[0, 1]$, and so the existence of negative estimates or estimates that are > 1 is a serious problem. We experimented with different schemes for normalising such values but found that none of them were successful.

2.1. Non-negative matrix factorisation (NNMF)

Non-negative matrix factorisation (NNMF, [9]) is an attractive alternative to PCA and SVD in cases where estimates are required to be positive e.g. if the estimates are word counts, as in [10]), or probabilities, as in this case. Of course, NNMF cannot guarantee that estimates will be ≤ 1 or that they will sum to 1 over a row of a confusion matrix, but the normalisations required to meet these conditions are less severe than those required to deal with negative estimates. In practice, we have found that the estimates produced by NNMF are always of the same order as the input values, which are themselves probabilities, and estimates > 1 are very rarely encountered.

It is not the intention to give a detailed account of NNMF in this paper: for a good introduction to NNMF, see, for example [9]. NNMF seeks to approximate an $n \times m$ non-negative matrix V

by the product of two non-negative matrices W and H :

$$V \approx WH. \quad (1)$$

W is a $n \times r$ matrix and H is a $r \times m$ matrix, where $r \leq \min(n, m)$. When $r < \min(n, m)$, the estimate of V , $\hat{V} = WH$, can be regarded as having been projected into and out of a lower-dimensional space r . The matrix V consists of m examples (the columns) of an n -dimensional random variable (the rows). It is thus apparent that the i 'th example, a column vector \mathbf{v}_i^T , is estimated as

$$\mathbf{v}_i^T \approx W\mathbf{h}_i^T, \quad (2)$$

where \mathbf{v}_i^T and \mathbf{h}_i^T are the corresponding columns of V and H . Hence each \mathbf{v}_i^T is estimated as a weighted sum of the columns of W , the j 'th column weighted by the scalar $\mathbf{h}(j)_i$. So the columns of W can be regarded as forming a set of (non-orthogonal) basis vectors that efficiently represent the structure of V . Estimation of W and V is accomplished by minimising a cost function between \hat{V} and V ; the distance function $D()$ used in the minimisation was the Frobenius norm i.e. $D(V, \hat{V}) = \sum_{i,j} (V_{i,j} - \hat{V}_{i,j})^2$. The minimisation algorithm used was that due to Lee and Seung [11].

2.2. Evidence for correlations within confusion matrices

The absolute value of the correlations between the elements of confusion matrices taken from 91 different speakers was estimated. Since there are $46 \times 46 = 2116$ elements in a confusion matrix (see section for details), there are $2116 \times 2116 = 4.4m$ such correlations. The top half of Figure 1 is a histogram of

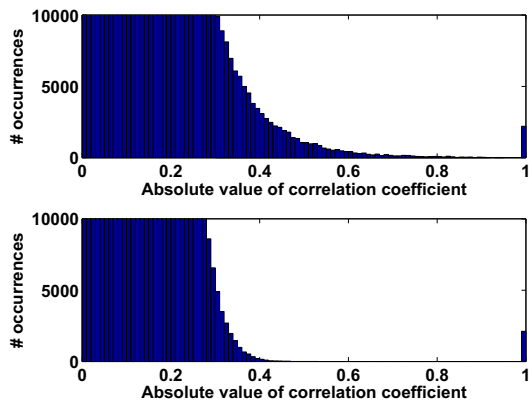


Figure 1: Top: Correlation of all elements of 46×46 confusion matrices from 91 speakers. Bottom: Comparison with randomly generated “confusion matrices”

these correlations: the y -axis has been clipped so that the small number of high correlations may be observed. For comparison, the bottom part of Figure 1 shows the distribution of the correlations taken from a synthetic dataset of the same size in which the confusion matrix elements had been randomly generated: there are no correlations above 0.48. The higher correlations were examined: many of them were phonetically plausible e.g. the confusion z/f has correlation 0.98 with the confusion s/f, the confusion v/sh has correlation 0.80 with the confusion zh/hh; some probably reflected recogniser insertions (which are also important to model); and some reflected statistical quirks that

were due to very low counts. Correlations between diagonal elements only and between elements in the same row were also measured before the experiments described in section 4 were undertaken and were also found to be significant. We were sufficiently encouraged by the presence of high correlations to proceed.

3. Speech Data and Confusion Matrices

3.1. Data

We conducted our experiments with a subset of the WSJCAM0 database [12]. The speech data used was parameterised to a 39-d vector consisting of 12 MFCCs + velocity + acceleration coefficients and log energy coefficients. The training-set consisted of 8246 utterances from the si_tr portion of WSJCAM0, a total of about 90 hours of speech from 91 speakers. It was used to train a set of 45 HMMs of monophones, each model being a three state, left-right model having a 25 component Gaussian mixture model associated with each state. A phonotactic bigram language model was estimated from the transcriptions of the same data. Using monophone models with a large number of mixture components per state was found to give better phone recognition performance than using triphone models with a smaller number of components. The test-set was the non-adaptation sentences of the si_dt set of WSJCAM0, consisting of 773 utterances from 20 different speakers.

3.2. Confusion matrices

Confusion matrices were estimated by alignment of the recogniser output to the phone transcription of the input text. In cases where multiple pronunciations were possible, only the most likely pronunciation was used. The ARPABET phoneme set (45 phonemes) was used and the confusion matrices were 46×46 matrices. Element i, j of a confusion matrix is the probability that the phoneme p_j is recognised when p_i is uttered ($1 \leq i, j \leq 45$), and is estimated in the standard way using the relative frequency of co-occurrence counts. Column 46 is the probability that phoneme p_i is inserted and row 46 the probability that phoneme p_j is deleted.

For each speaker from both the training- and test-sets, the “true” confusion matrix is defined as the confusion matrix estimated using all the available utterances from the speaker: there were an average of 100 utterances per speaker and over 7000 phonemes per speaker, giving an average of over 153 phonemes per row, although in practice, the values for each row varied widely. In addition to the “true” confusion matrix, “partial” confusion matrices for each speaker were computed from randomly selected subsets of utterances: 5, 8, 12, 18, 25 and 50 utterances were used. The “true” confusion matrix for training-set speaker S_i is denoted as CM^i and an estimate of it is \widehat{CM}^i . The confusion matrix made using U utterances is denoted as CM_U^i . The mean confusion matrix averaged over all the training-set speakers, is denoted \overline{CM} .

4. Models Used and Experimental Procedure

We experimented with two models for estimating confusion matrices:

1. **Direct.** Each column of V is a complete confusion matrix (written out column by column) from a training-set speaker. To estimate a confusion matrix from a partial

matrix CM_U^i , the partial matrix is added as an extra column to V , NNMF is applied to V , and the resulting estimate \widehat{CM}^i is retrieved from \widehat{V} . The process is iterated until the estimate obtained converges.

2. **DiagRow**. A potential problem with **Direct** is the high dimensionality of the elements to be estimated (2116) compared with the small number of samples (91). We therefore sought an approach that was more focussed on sections of the confusion matrices that were liable to be well-correlated and hence have good predictive properties. Our premise in **DiagRow** is that diagonal elements of confusion matrices exhibit good correlation, as does any single row. The first assertion could be put informally as “low (or high) accuracy on phoneme A implies low (or high) accuracy on phoneme B .”, and the second as “phoneme C tends typically to be confused with phonemes $D, E \dots$ ”. The relevant correlations are all contained within the analysis described in section 2.2, but the proportion of higher correlations across diagonal elements and across single rows may be higher than when the complete matrix is considered. The algorithm for estimation using **DiagRow** iterates estimation of the diagonal elements followed by estimation of the row elements of the confusion matrices of each speaker. In pseudo-code:

```

Load the diagonal entries from speaker  $S_i$  into the  $i$ 'th
column of  $V_{diag}$ .
 $Current \leftarrow CM_U^i$ ;  $Last = RAND$ 
while  $\mathcal{D}(Current, Last) > \epsilon$  do
  Update diagonal entries
  for Each speaker  $S_i$  do
    Load the diagonal entries of  $Current$  into the
    last column of  $V_{diag}$  and do NNMF on  $V_{diag}$ .
    Replace the diagonal entries of  $Current$  with
    the last column of  $\widehat{V}_{diag}$ 
  end for
  Update row entries
  for Each speaker  $S_i$  do
    for Each row  $CM_U^i(j)$  do
      Load the  $j$ 'th row from training-set speaker
       $S_i$  into the  $i$ 'th column of  $V_{col}$  and do NNMF
      on  $V_{col}$ .
      Replace the  $j$ 'th row of current with the last
      column of  $\widehat{V}_{col}$ 
    end for
  end for
end while
 $\widehat{CM}^i = Last$ 
Re-normalise rows of  $\widehat{CM}^i$  to sum to 1.0

```

4.1. Smoothing and Masking

When a small number of utterances, U , has been used to estimate CM_U^i , some rows of CM_U^i may have a very small number of co-occurrence counts from which to estimate the confusion matrix probabilities, or even have zero counts. For instance, when five utterances are used to estimate speakers' confusion matrices, 10.9% of the rows have no co-occurrence counts and nearly half (43.9%) have fewer than five counts. It was therefore important to apply some smoothing to the partial confusion matrices. This was done in a simple way by choosing a count threshold CT , and then replacing any partial confusion matrix row that had been estimated from fewer than CT counts with

the same row of the mean confusion matrix, \overline{CM} . We also observed that the “true” confusion matrices are very sparse: the mean number of zeros in a training-set speaker's confusion matrix is 1530.2 (out of a total of 2116 elements), with a standard deviation of only 58. However, only 145 of the entries in the mean confusion matrix \overline{CM} are exactly zero, although 1999 are ≤ 0.01 . We experimented with removing from the estimation technique any element whose equivalent entry in \overline{CM} was less than a “masking” threshold MT . This gives a very large reduction in dimensionality and hence in processing time.

5. Results

We use a weighted distance squared difference (WSD) measure \mathcal{D} to assess the quality of estimates of speakers' confusion matrices:

$$\mathcal{D}(CM^i, \widehat{CM}^i) = \frac{1}{N} \sum_{i=1}^N \sum_j \Pr(j) \sum_k (CM^i(j, k) - \widehat{CM}^i(j, k))^2, \quad (3)$$

where N is the number of speakers. In words, the squared difference between elements (j, k) of the “true” confusion matrix for speaker S_i , CM^i , and its estimate, \widehat{CM}^i , is weighted by the probability that phoneme p_j is spoken. This gives a more realistic guide to the quality of the estimates than using an unweighted distance, as $\Pr(p_j)$ varies considerably, some phonemes being rarely spoken.

Results are the best obtained after varying the number of dimensions r (as explained in section 2.1) and the thresholds CT and MT described in section 4.1. For the **Direct** technique, the optimum value of r was 100–200 for the training-set, and about 50 for the test-set. All techniques benefited from some smoothing with the mean confusion matrix, and the best values of CT were in the range 3–5. Removing elements that were (on average) close to zero by setting $MT > 0$ always worsened results, which implies that the NNMF process takes advantage of the full structure of confusion matrices from different speakers.

Figure 2 shows the results on the training-set. The dashed line shows the WSD between the supplied partial confusion matrices (over all training-set speakers), and the solid line the WSD to estimates made from either the **Direct** or **RowDiag** techniques: the WSDs given by these two techniques were within 0.01 of each other. Note that an estimate is useful only when it is closer to the true confusion matrix than both the supplied partial matrix and the mean confusion matrix, \overline{CM} ; if \overline{CM} is closer than the estimated matrix to the true matrix, this could be used in preference to the estimated matrix. Hence the horizontal line on the plot shows the mean WSD between the true confusion matrices and the mean confusion matrix. For every value of U (the number of utterances supplied to make the partial confusion matrix), the techniques give estimates of confusion matrices that have lower WSDs to the true confusion matrices than the partial confusion matrices: *much* closer when U is small. When only five utterances per speaker are used for confusion matrix estimation, the WSD between the true and estimated matrices is the same as the WSD between the true and mean confusion matrix, so there is no advantage in using these estimates. When 50 utterances are used, the supplied partial confusion matrix for a speaker is already quite close to the true matrix, and the estimates do not improve the matrices much. In between these extremes, the techniques show improved estimates. For instance, when $U = 12$, the WSD obtained is equivalent to a supplied partial matrix with $U \approx 25$, so the technique effectively doubles the

number of utterances available for confusion estimation.

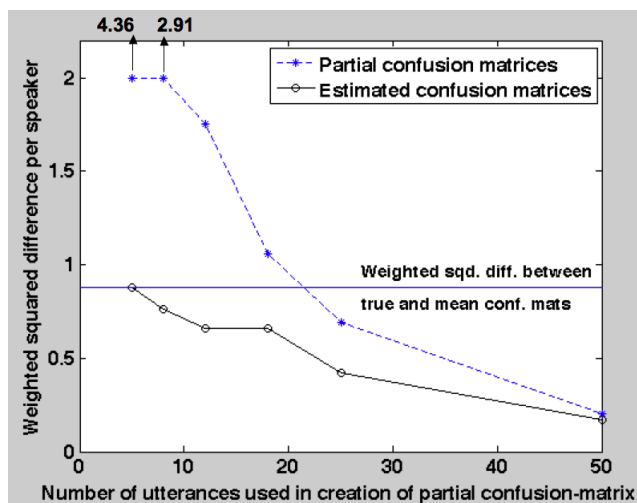


Figure 2: Training set results

Figure 3 shows the results on the test-set. Note that the maximum number of “partial” utterances available for the test-set was only 25, rather than 50 as in the training-set experiments. On the test-set, results from **Direct** and **RowDiag** differed, and so both are plotted here. The absolute values of WSD obtained for different values of U are similar to those obtained on the training-set, but because the WSD between the “true” and mean confusion matrices is higher for the test-set speakers (since the mean confusion matrix is estimated from the training-set speakers only), the resulting gain is greater, and is significant for only five utterances. The figure indicates that the WSD of the matrices estimated using five utterances of supplied data is equivalent to matrices directly estimated from about 18 utterances. Time did not permit us to investigate whether any gain could be obtained for fewer than five utterances.

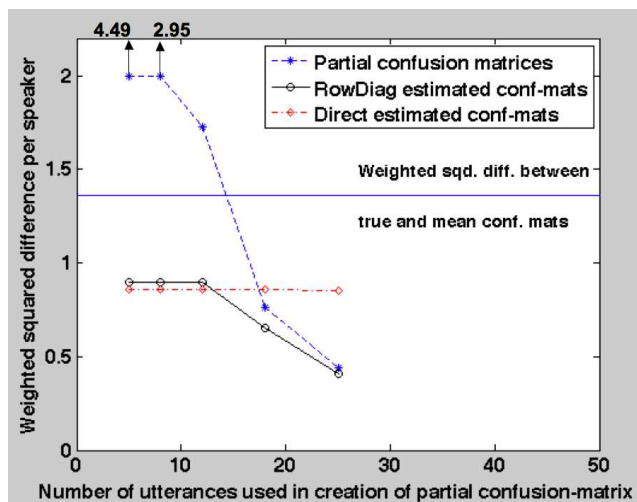


Figure 3: Test set results

6. Discussion and Future Work

We have described a technique for making improved estimates of an individual speaker’s confusion matrix given an estimate that has been derived from a small set of utterances from the

speaker. The technique uses non-negative matrix factorisation to find structure within confusion matrices, and this structure is exploited to make improved estimates. Our results indicate that when the data is very sparse, although our techniques give estimates that are much better than the estimates made directly from the supplied data, these estimates are no better in terms of distance from the speaker’s “true” confusion matrix than using the mean confusion matrix. However, when the data supplied is adequate, we obtain estimates that are equivalent to matrices made from many more than the utterances actually available from the speaker.

Another approach to the problem is to suppose that a confusion matrix is made up of two components: a component that is due to the individual speaker and a component that is due to the properties of the speech recogniser. This leads to the idea of “speaker factors” that interact with parameters due to the recogniser. This forms the basis of ongoing research, as does the use of the estimated confusion matrices within a recogniser to improve accuracy.

7. References

- [1] S. Srinivasan and D. Petkovic. Phonetic confusion matrix based spoken document retrieval. In *Proc. 23rd annual international ACM SIGIR conference on Research and Development In Information Retrieval*, pages 81–87. ACM Press New York, NY, USA, 2000.
- [2] M. Levit, H. Alshawi, A. Gorin, and E. Nöth. Context-Sensitive Evaluation and Correction of Phone Recognition Output. In *Proc. Eighth European Conference on Speech Communication and Technology*. ISCA, 2003.
- [3] O. Cabellero-Morales and S.J. Cox. Modelling confusion-matrices to improve speech recognition accuracy, with an application to dysarthric speech. In *Proc. 10th International Conference on Spoken Language Processing (Interspeech)*, Antwerp, August 2007.
- [4] J.B. Allen. *Articulation and Intelligibility*. Synthesis Lectures on Speech and Audio Processing. Morgan and Claypool, 2005.
- [5] R. Kuhn, P. Nguyen, J.C. Junqua, A. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini. Eigenvoices for speaker adaptation. In *Proc. Fifth International Conference on Spoken Language Processing*, 1998.
- [6] A.H. Abolhassani, S.A. Selouani, and D. O’Shaughnessy. Speech enhancement using PCA and variance of the reconstruction error in distributed speech recognition. In *Proc. IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 19–23, 2007.
- [7] J.R. Bellegarda. A multispan language modeling framework for large vocabulary speech recognition. *IEEE Transactions on Speech and Audio Processing*, 6(5):456–467, September 1998.
- [8] L. Reveret and C. Benoit. A New 3D Lip Model For Analysis and Synthesis Of Lip Motion In Speech Production. In *Proc. of International Conference on Auditory-Visual Speech Processing (AVSP)*. ISCA, 1998.
- [9] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.
- [10] M. Novak and R. Mammone. Use of non-negative matrix factorization for language modeladaptation in a lecture transcription task. In *Proc. IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, 2001.
- [11] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, volume 13, pages 556–562. Morgan-Kaufmann, 2001.
- [12] T. Robinson et al. WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, pages 81–84, 1995.