INTERSPEECH 2009 BRIGHTON

On the Estimation and the Use of Confusion-Matrices for Improving ASR Accuracy

Omar Caballero Morales and Stephen Cox

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

S.Caballero-morales@uea.ac.uk, s.j.cox@uea.ac.uk

10.21437/Interspeech.2009-207

Abstract

In previous work, we described how learning the pattern of recognition errors made by an individual using a certain ASR system leads to increased recognition accuracy compared with a standard MLLR adaptation approach. This was the case for low-intelligibility speakers with dysarthric speech, but no improvement was observed for normal speakers. In this paper, we describe an alternative method for obtaining the training data for confusion-matrix estimation for normal speakers which is more effective than our previous technique. We also address the issue of data sparsity in estimation of confusion-matrices by using non-negative matrix factorization (NMF) to discover structure within them. The confusion-matrix estimates made using these techniques are integrated into the ASR process using a technique termed as "metamodels", and the results presented here show statistically significant gains in word recognition accuracy when applied to normal speech.

Index Terms: confusion-matrix modelling, metamodels, non-negative matrix factorization

1. Introduction

In previous work on improving the recognition of disordered (dysarthric) speech, we argued that conventional speaker adaptation techniques were inappropriate and that learning how to correct phone sequences is a better strategy. We achieved this by modelling a speaker's pattern of errors, represented in a phoneme confusion-matrix, in two different ways: by using a set of discrete Hidden Markov Models (HMMs), which we termed "metamodels" [1], and a network of weighted finite state transducers (WFSTs) at the confusion-matrix, lexicon, and language model levels [2]. Although these techniques significantly increased word recognition accuracy for low intelligibility speakers, they failed to increase performance with high intelligibility dysarthric speakers, or with normal speakers, and in some cases, the performance was worse than the baseline [1]. We observed in later experiments that the performance of the metamodels and the WFSTs relied critically on the accuracy of the phoneme output sequences used for confusion-matrix estimation. This observation led to the first issue discussed in this paper, which is a method to obtain training phoneme sequences that increase performance when used with metamodels. Details of the method are given in Section 2.1, and the metamodels are reviewed in Section 2.3.

The second issue is related to the problem observed when the data available for confusion-matrix estimation is small, which leads to poor estimates, and in practice, this is the normal situation. A novel approach presented in [3] made use of Non-negative Matrix Factorization (NMF) to find structure within confusion-matrices from a set of "training" speakers, which could be used to make improved estimates for a "test" speaker given only a few samples from his/her speech. An advantage of this technique is that it was able to remove some of the noise present in the sparse estimates, while retaining the particular speaker's confusion-matrix patterns. This approach is reviewed in Section 2.2.

Here, the NMF technique is extended to estimate also insertion patterns, and we analyse the effect of pre-processing adjustments (e.g. smoothing) in the NMF estimates. The quality of the estimates is dependent on two factors, the degree of smoothing and the NMF process itself, and here we measure the effect of each factor. This is presented in Section 4.1. In addition, we integrate the NMF estimates into the metamodels, and evaluate their performance in word recognition accuracy for normal speakers.

The results presented in Section 4.2 show that the proposed method to obtain the phoneme sequences for confusion-matrix estimation statistically improved the performance of the metamodels for normal speakers when compared with the baseline, which was MLLR adaptation. The use of NMF also increased the previous performance of the metamodels when few utterances were available for confusion-matrix estimation. In Section 5 we discuss these findings.

2. Preliminaries

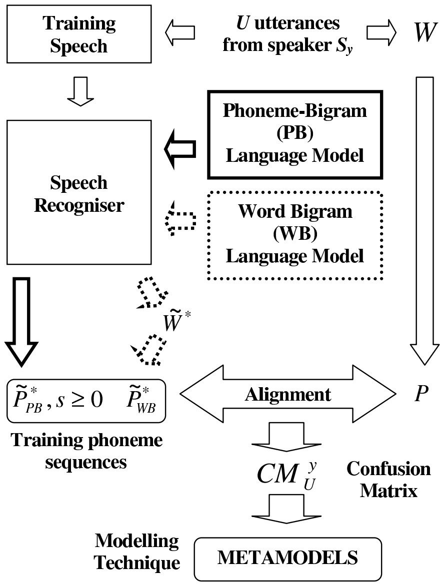
2.1. Confusion-Matrix Estimation

The procedure to estimate a confusion-matrix is illustrated in Figure 1. We define W as the sequence of words that the speaker wished to utter, and $\tilde{P}$ as the sequence of phonemes decoded by an ASR system: hence, $\tilde{P}$ includes effects of both mispronunciations by a speaker and errors made by the recogniser. If W is transcribed at the phonemic level as P, and a sub-set $\tilde{P}^* \in \tilde{P}$ is selected for "training", a confusion-matrix that models $Pr(\tilde{p}_i{}^*|p_i)$ can be estimated from the alignment of $\{P, \tilde{P}^*\}$[1]. Because p_i is the i'th phoneme in the postulated phoneme sequence P, and $\tilde{p}_i^*$ the i'th phoneme in the decoded sequence $\tilde{P}^*$, $Pr(\tilde{p}_i{}^*|p_i)$ represents the probability that the phoneme p_i^* is recognised when p_i is uttered. The alignment also identifies phonemes $\tilde{p}_i^*$ that are inserted, and phonemes p_i that are deleted.

Now we take some nomenclature from [3] to define confusion-matrices estimated from large amounts of data (**target**) and from sparse data (**partial**). For a speaker S_y, a **target** confusion-matrix, which is estimated using all the available utterances from that speaker, is designated as CM^y. In addition, **partial** confusion-matrices from that speaker, defined as CM_U^y, are es-

[1] $\tilde{P}^*$ can be obtained by using a phoneme language model $\left(\tilde{P}^*_{PB}\right)$, or a word language model $\left(\tilde{P}^*_{WB}\right)$. Details are given in Section 2.1.1.

Figure 1: *Procedure for confusion-matrix estimation.*



timated by using $U = \{5, 10, 15, 20, 30\}$ utterances. In Section 2.1.1, we present the method to obtain the most suitable $\tilde{P}^*$ for confusion-matrix estimation. Details of the application of NMF to the partial confusion-matrices, and how they are integrated into the recognition process, are presented in the next sections.

2.1.1. A Method for Obtaining $\tilde{P}^*$

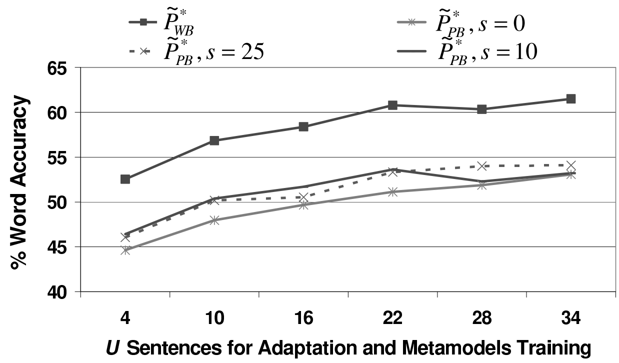
In previous experiments [1], $\tilde{P}^*$ was decoded by using a phoneme-loop language model. This is equivalent to using a phoneme-bigram (PB) language model with a **grammar scale factor**[2] of zero. These phoneme sequences can be expressed as $\{\tilde{P}^*_{PB}, s = 0\}$, where *PB* denotes the kind of language model used, and *s* the magnitude of the grammar scale factor.

In later experiments with disordered speech, the grammar scale factor was increased (e.g., $s = 10$, and $s = 25$). As shown in Figure 2, this change improved the metamodels' performance. An alternative method is shown as the dotted lines in Figure 1. Here, the training speech is decoded by using a word-bigram (WB) language model with a fixed grammar scale factor ($s_{WB} = 30$). The phonemic transcriptions of the words decoded then provide the phoneme sequences $\tilde{P}^*_{WB}$. Because of the extra constraint of a word-level language model in the decoding process, prior to conversion to phonemes, the accuracy of $\tilde{P}^*_{WB}$ is higher than just using a "phoneme loop" decoder. As shown in Figure 2, the metamodels built using these sequences for training-data achieved the highest word recognition accuracy (Section 2.3 describes in detail how the metamodels are used in the recognition process).

It is interesting to note that, while $\tilde{P}^*_{WB}$ had the highest phoneme accuracy, and $\{\tilde{P}^*_{PB}, s = 0\}$ the lowest, this did not necessarily correlate with the performance of the metamodels. The key point is the rate of phoneme deletions. Analysis showed that $\tilde{P}^*_{WB}$ could be quite accurate, with a low number of insertion and substitution errors, but could also have a high rate of deletion errors (e.g., due to deleted phoneme sequences), which decreases the performance of the modelling technique [2]. In this case, metamodels trained with $\{\tilde{P}^*_{PB}, s \geq 0\}$ performed better for some speakers with disordered speech. Hence, the use of $\tilde{P}^*$ was based on two measures: high phoneme accuracy (from the training set), and low rate of deletions. For normal speakers, $\tilde{P}^* = \tilde{P}^*_{WB}$ was a suitable choice.

[2]The grammar scale factor controls the influence of the language model on the decoding process. As the grammar scale factor increases, the decoding relies more on the language model than the acoustic signal.

Figure 2: *Performance of the metamodels trained with different sets of phoneme strings $\tilde{P}^*$: mean word recognition accuracy across all the speakers from the Nemours Database of Dysarthric Speech [4].*



2.2. Non-negative Matrix Factorization (NMF)

In [3], we reported that there was inherent structure and correlations present in the confusion-matrices estimated from data from speakers in the Wall Street Journal (WSJ) database. NMF was proposed as an approach to utilise these correlations to estimate an individual speaker's confusion-matrix given some sparse data from him/her. An advantage of using NMF over other similar methods, such as Principal Component Analysis (PCA) and Singular Value Decomposition (SVD), is the non-negativity property. This makes it suitable to estimate confusion-matrix probabilities (e.g., $Pr(\tilde{p}_i{}^*|p_i)$) that are restricted to the range $[0, 1]$. Although there is no guarantee that the estimates will be ≤ 1, the normalisations required are less severe than those required if negative estimates are present.

NMF seeks to approximate an $n \times m$ non-negative matrix V by the product of two non-negative matrices W and H:

$$V \approx WH. \tag{1}$$

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 being projected into and out of a lower-dimensional space r [5]. The columns of W are regarded as forming a set of (non-orthogonal) basis vectors that efficiently represent the structure of V, with the columns of H acting as weights for individual column vectors of V [3].

Estimation of V is accomplished by minimising a distance function between WH and V, which is defined by the Frobenius norm [5]. The minimisation algorithm is the one proposed by Lee and Seung in [6].

2.2.1. Target Confusion-Matrix Estimation

In this work, the **Direct** model as defined in [3] was used for confusion-matrix estimation. Each column of V is a **target** confusion-matrix CM^x (written out column by column) from a training-set speaker S_x. To estimate a **target** confusion-matrix CM^y from a **partial** confusion-matrix CM^y_U of a "test" speaker S_y, CM^y_U is added as an extra column to V. When NMF is applied to V, the estimated confusion-matrix $\widehat{CM}^y$ is retrieved from $\hat{V}$ and is re-normalised so that its rows sum to 1.0. As presented in [3], a weighted distance squared difference measure $D(CM^y, \widehat{CM}^y)$ was used to assess the quality of the estimates of $\widehat{CM}^y$. The process is iterated until the obtained estimates converge. More information of the algorithm can be found in [3].

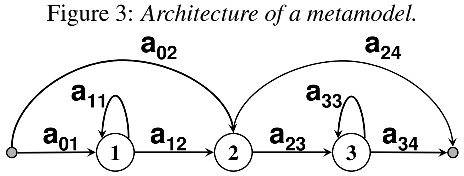
2.2.2. Smoothing

If the data available from a speaker is very small, the **partial** confusion-matrix CM_U^y will be too sparse to make an improved estimate using NMF. Hence, CM_U^y is smoothed using a speaker-independent confusion-matrix $\overline{CM}$ that is well estimated from the training data. If the total number of non-zero elements in CM_U^y is less than a threshold $CT = \{0, 2, 5\}$, the row is replaced by the equivalent row of $\overline{CM}$.

2.3. Metamodels

The use of "metamodels" as a technique to integrate $Pr(\tilde{p}_i{}^*|p_i)$ into the word recognition process was first described in [1] and was also used in [7]. The architecture of the metamodel of a phoneme is shown in Figure 3. Each state of a metamodel has a discrete probability distribution over the symbols for the set of phonemes, plus an additional symbol labelled DELETION. The central state (2) of a metamodel for a certain phoneme models correct decodings, substitutions and deletions of this phoneme made by the recogniser. States 1 and 3 model (possibly multiple) insertions before and after the phoneme. The parameters of the metamodels are trained by using the partial confusion-matrices CM_U^y estimated from accurate alignments of P and $\tilde{P}^*$ as detailed in section 2.1.

Figure 3: *Architecture of a metamodel.*



As an example of the operation of a metamodel, consider a hypothetical phoneme that is always decoded correctly without substitutions, deletions or insertions. In this case, the discrete distribution associated with the central state would consist of zeros except for the probability associated with the symbol for the phoneme itself, which would be 1.0. In addition, the transition probabilities a_{02} and a_{24} would be set to 1.0 so that no insertions could be made. Before recognition, a language model is used to compile a "meta-recogniser" network, which is identical to the network used in a standard word recogniser except that the nodes of the network are the appropriate metamodels rather than the acoustic models used by the word recogniser. At recognition time, the output phoneme sequence $\tilde{P}^*$ is passed to the meta-recogniser to produce a set of word hypotheses [1].

In [3], NMF was used to obtain improved estimates of confusion-matrices (as measured by the Frobenius Norm between the estimated and target matrices), but these were not used to improve ASR. Here, we used NMF to estimate the set of discrete probabilities associated with the states 1, 2 and 3 of the metamodel of Figure 3. Note that the set associated with state 2 is the same as a row of a confusion-matrix, and the sets associated with states 1 and 3 represent the probabilities of insertion of phonemes before and after the phoneme, respectively. For the states 1 and 3, the transitions a_{11} and a_{33} were set to zero for simplicity. Note that all transition probabilities are estimated by counting of phoneme occurrences.

3. Speech Data and Baseline Recogniser

The Wall Street Journal (WSJ) database was used to build the baseline speech recogniser. The training set consisted of the WSJ data from 92 speakers in set *si_tr*. This was used to construct 45 monophone acoustic models. The models were standard three-state left-to-right topology with eight mixture components per state. The front-end used 12 MFCCs plus energy plus delta and acceleration coefficients.

For the NMF experiments, the training-speakers S_x for V (see Section 2.2.1) consisted of 85 speakers from the *si_tr* set, which were also used to estimate $\overline{CM}$. 10 test-speakers S_y for V were selected from the set *si_dt* of the same database. Note that from the training-speakers, only **target** confusion-matrices CM^x were estimated. For the test-speakers, **partial** CM_U^y and **target** CM^y confusion-matrices were estimated in order to evaluate the quality of the NMF estimates of $\widehat{CM}^y$.

Supervised Maximum Likelihood Linear Regression (MLLR) adaptation [8] was implemented using the same sets of utterances U selected for confusion-matrix estimation. A regression class tree with 32 terminal nodes was used for this purpose. As shown in Table 1, the mean number of MLLR transformations increased as more utterances were used. The adapted acoustic models represent the baseline for our experiments.

Table 1: *Mean number of MLLR transformations across all test speakers using different sets of adaptation data.*

Adaptation Data (U)	5	10	15	20	30
Mean Transformations	4	7	10	11	12

A word-bigram language model, estimated from the data of the *si_tr* speakers, was used to obtain $\tilde{P}^*_{WB}$ and estimate CM^x with the unadapted baseline. In order to keep these sequences independent from those of the test-set speakers, the word-bigram language model used to decode $\tilde{P}^*_{WB}$ for CM^y and CM_U^y, was estimated from the data of the selected test-speakers of the *si_dt* set. In all cases, a grammar scale factor s_{WB} of 10 was used. The metamodels were tested using all the utterances available from the speakers S_y.

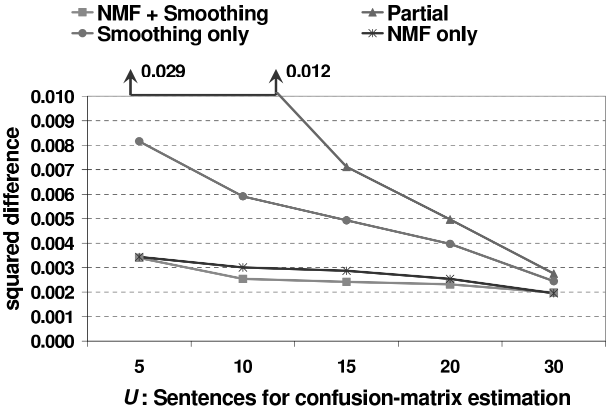
4. Results

4.1. Effect of Smoothing on the NMF Estimates

We were interested to see whether the improvements in the estimates of confusion-matrices reported in gain in [3] came merely from the simple smoothing procedure rather than the NMF. To analyse these factors separately, the estimates of CM^y ($\widehat{CM}^y$) were obtained by using (1) the smoothing only, (2) the NMF procedure only, and (3) the NMF procedure plus the smoothing as described in [3]. Figure 4 shows the squared difference $D(CM^y, \widehat{CM}^y)$ computed for the confusion-matrices estimated as described above. In addition, the difference obtained when CM_U^y is used as an estimate of CM^y is plotted. The minimum error for the test speakers using the NMF procedure was obtained when r varied within the range $[10-30]$.

Although the smoothing-only estimate is always better than the estimates made only from partial data, it is much worse than both the NMF-only and the NMF + Smoothing estimates when low number of adaptation utterances are used. When enough data is available for estimation (e.g., $U = 30$), all the estimates converge to approximately the same error.

Figure 4: *Weighted squared difference $D(CM^y, \widehat{CM}^y)$ - adapted baseline.*



A factorial analysis was made to determine how the NMF and smoothing interacted. Two factors were considered: (1) the smoothing ("1" if it is implemented, "0" otherwise), and (2) the NMF estimation process ("1" if it is implemented, "0" otherwise). Table 2 shows the scaled mean squared difference across all test-speakers for each set of conditions. When Smoothing = 0 and NMF = 0, only the mean **partial** estimates are considered. When no smoothing is done, the NMF-only estimates show a mean difference of 0.276. However when the smoothing is included, this difference only falls to just 0.253. Smoothing, without NMF, reaches 0.508, which is almost two times the difference of the NMF-only estimates. The effect of the smoothing is a reduction of the NMF difference by only 0.276 - 0.253 = 0.023, which we conclude is probably not significant.

Table 2: *Mean squared difference across all test-speakers and all U sets under conditions for NMF and smoothing. Mean value* $\times 100$

	NMF=0	NMF=1
Smoothing = 0	1.100	0.276
Smoothing = 1	0.508	0.253

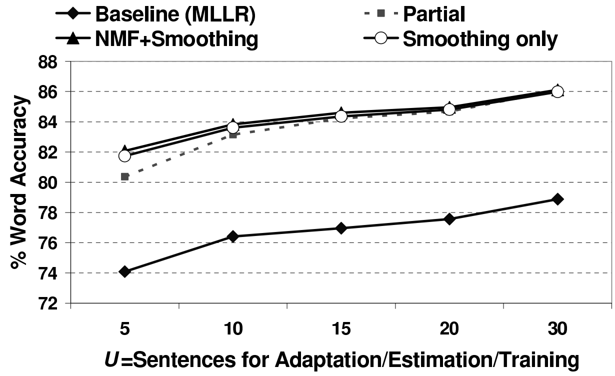
4.2. NMF Estimates on the Metamodels

Figure 5 shows the performance of the metamodels using MLLR adapted acoustic models. The metamodels trained with **partial** estimates improved over the adapted baseline, something that was not achieved on "normal" speech in previous work. When the NMF+Smoothing estimates were used, the accuracy of the metamodels improved when the training data was small. These improvements were statistically significant when five and ten utterances were used for training/estimation. The metamodels trained with the smoothing-only estimates, although not better, produced a very similar improvement.

5. Discussion and Future Work

In this work, we have consolidated our work on the use of error modelling to increase recognition accuracy. Our previous work showed that the "metamodels" approach was highly effective for improving recognition accuracy on dysarthric speech, and this work demonstrates that it can also outperform a standard adaptation technique such as MLLR on normal speakers when a more sophisticated training procedure is used for training the metamodels. A key point for future research is to refine the

Figure 5: *Mean word recognition accuracy of the metamodels across all test-speakers - adapted baseline.*



meta-modelling to handle deletion errors better, as these can affect the performance of the metamodels severely.

We have also investigated the use of non-negative matrix factorisation (NMF) and simple smoothing to obtain improved estimates of a speaker's confusion-matrix when only sparse training data is available. We have found that this approach leads to small but statistically significantly increases of performance of our technique when the amount of data available for estimation is low (five and ten utterances). Although NMF produced better estimates of the confusion matrices than those obtained using a simple smoothing technique, the recognition performance from metamodels trained on data using the two techniques was not statistically significantly different. Our next step is to attempt to integrate the sparse estimation techniques described here with the use of weighted finite state transducers (WFSTs) for error correction, and to explore alternative smoothing techniques. Measure of improvement when using other adaptation techniques (e.g., MAP) will also be studied.

6. References

[1] Caballero, Omar and Cox, Stephen, "Modelling confusion matrices to improve speech recognition accuracy, with an application to dysarthric speech," *Interspeech 2007*, 2007.

[2] ——, "Application of weighted finite-state transducers to improve recognition accuracy for dysarthric speech," *Interspeech 2008*, 2008.

[3] Cox, Stephen, "On estimation of a speakers confusion matrix from sparse data," *Interspeech 2008*, 2008.

[4] Bunnel, H.T. and Polikoff, J.B., "The nemours database of dysarthric speech," *Proceedings of ICSLP*, 1996.

[5] Lee, D.D. and Seung, H.S., "Learning the parts of objects by nonnegative matrix factorization," *Nature*, 1999.

[6] ——, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[7] Matsumasa, Hironori, Takiguchi, Tetsuya, Ariki, Yasuo, LI, Ichao, and Nakabayashi, Toshitaka, "Integration of metamodel and acoustic model for speech recognition," *Interspeech 2008*, 2008.

[8] Leggetter, C.J. and Woodland, P.C., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–85, 1995.