ADAPTATION OF FREQUENCY BAND INFLUENCE FOR NON-NATIVE SPEECH RECOGNITION

Martin Raab[1,2]*, Tobias Herbig*[1,3]*, Raymond Brueckner*[1]*, Rainer Gruhn*[1,3]*, Elmar Nöth*[2]

[1]Harman Becker Automotive Systems (HBAS), Speech Dialog Systems, Ulm, Germany
[2]University of Erlangen Nuremberg, Dept. of Pattern Recognition, Erlangen, Germany
[3]University of Ulm, Dept. of Information Technology, Ulm, Germany
mraab@harmanbecker.com

ABSTRACT

For voice controlled car navigation systems, multilinguality is a big challenge. The goals are clear. Users drive to other countries and need to enter foreign city names, at the same time it is likely that they will keep interacting in their native language for other commands. One important aspect is that the utterances the users produce differ from native speaker utterances, they have a non-native accent.

The motivation for our work is that people hear better at low frequencies and know that low frequencies are more important for producing understandable utterances in the foreign language. Therefore they first aim to copy the low frequency behavior of the foreign language. Additionally, changes in mid to high frequencies are caused by little tongue movements. These subtle changes are hard to control for non-native speakers. Together both reasons cause the effect that non-native speech differs stronger from native speech for mid-range frequencies.

Thus we analyze if speech recognition of non-native speakers can be improved by lowering the influence of mid to high frequencies. We achieve this through increasing some variances of the Gaussians. This leads to an reduced influence of differences in the corresponding frequency band on the likelihood output of a Gaussian. This way we can model the selective mismatch between native training data and non-native test data.

1 Introduction

Non-native speech is a severe problem for all system which have to cope with multilingual input. Applications that face this problem are personal navigation devices and automated travel booking operators. For computer aided language learning systems and possible speech operated MP3 players the problem is even more severe, as they have to deal with non-native speech most of the time.

Many publications can report significant improvements with non-native training or adaptation data [2, 3, 4]. Fewer works can achieve improvements without non-native data [7, 8, 10]. Making the distinction between using or not using non-native data is fundamental, as approaches of the first kind would need n^2 accent data bases to cover all possible accents of n languages. Although a recent overview of some of the authors [6] presents many accent databases, applying these approaches to a commercial product delivered in many languages would require a tremendous effort.

In Arslan's PhD thesis [1] it was shown that mid-range frequencies (1kHz-2.5 kHz) are more important for accent detection than the frequencies below 1 kHz. As mentioned, there are two

reasons for this. The preciseness of tongue movements and the perceptional system of humans. In the original work this effect was used to improve accent detection rates.

Our goal is to improve speech recognition performance for non-native speech. As stated above, non-native speech differs more severely for higher frequencies. It should be possible to improve speech recognition performance for non-native speech by increasing the weight on lower frequencies. We achieve this goal through modifications of the covariance matrices of the Gaussians. This way errors in higher frequencies have less impact on the likelihood that a feature vector was generated by this Gaussian. We call this technique Frequency Band Weight Adaptation (FBWA).

In order to perform such an operation, the transformations that are typically applied to a speech signal have to be considered. After the transformations (Linear Discriminant Analysis [LDA], Discrete Cosine Transform [DCT]) it is no longer straightforward to decide how a Gaussian has to be changed to treat different speech frequencies differently. Therefore the transformations have to be undone before the Gaussians are modified, and to be reapplied after that.

An advantage of this approach is that an existing recognizer can be modified very fast once there are a set of weights determined for better non-native speech recognition. At the same time, the modification can be undone for the recognition of native speech and no non-native data is needed to perform the modification.

The remainder of this paper is organized as follows. Section 2 describes the the algorithm we apply to the Gaussians. The experimental setup is given in Section 3. In Section 4 we present our results. A conclusion is drawn in Section 5.

2 Frequency Band Weight Adaptation

2.1 Modification of a Gaussian

The first question to answer is how a multivariate Gaussian has to be modified to weight some dimensions higher than others for its likelihood output. For this general aspect, it is irrelevant what the single dimensions of the Gaussian mean. This will be discussed in the next part.

A multivariate Gaussian A in a n-dimensional feature space is defined as

$$A(x) = \frac{1}{(2\pi)^{n/2}\,|\Sigma|^{1/2}}\, e^{-1/2(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

with the covariance matrix Σ

$$\Sigma = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nn} \end{pmatrix}$$

The term in front of the Euler constant is independent of the feature vector x and only necessary to normalize the overall Gaussian probability to 1. To achieve our desired modification only the exponent, which is the Mahalanobis distance has to be considered. The front term will automatically normalize the overall Gaussian probability to one.

The smaller the Mahalanobis distance between the feature vector and the mean vector is, the higher the likelihood of the Gaussian. Differences in one dimension will have less impact on the n-dimensional Mahalanobis distance, when the variance of the Gaussian is high in this dimension. Thus, if we modify the variance in one dimension i with a factor $(g_i)^2$, we can change the likelihood calculation as desired.

If the variance of each dimension is multiplied with a factor $(g_i)^2$, the covariances v_{ij} are also affected. The new covariances $\overline{v_{ij}}$ are

$$\overline{v_{ij}} = \sqrt{\overline{v_{ii}}} * \sqrt{\overline{v_{jj}}} = \sqrt{g_i^2 v_{ii}} * \sqrt{g_j^2 v_{jj}} = g_i g_j * \sqrt{v_{ii} v_{jj}} = g_i g_j v_{ij} \; g_i, g_j > 0$$

This leads to the new covariance matrix $\overline{\Sigma}$

$$\overline{\Sigma} = \begin{pmatrix} g_1^2 v_{11} & g_1 g_2 v_{12} & \cdots & g_1 g_n v_{1n} \\ g_2 g_1 v_{21} & g_2^2 v_{22} & \cdots & g_2 g_n v_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ g_n g_1 v_{n1} & g_n g_2 v_{n2} & \cdots & g_n^2 v_{nn} \end{pmatrix}$$

which is the same as Σ multiplied with $G * \Sigma * G^T$ where G is defined as

$$G = \begin{pmatrix} g_1 & 0 & \ldots & 0 \\ 0 & g_2 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & g_n \end{pmatrix}$$

2.2 Transformation of a Gaussian

Both an LDA and a DCT are applied to the frequency representation of the speech signal, before the Gaussians are estimated. Thus the dimensions on which the codebook Gaussians are based are not directly related to different frequencies in the speech signal.

The LDA and the DCT are matrix multiplications, and thus linear transformations. A linear transformation is also an affine transformation, with a translation part of zero. Gaussians remain Gaussians under both of them. More specifically, a Gaussian A as defined above becomes the Gaussian B under the linear transformation matrix T. The new Gaussian B has the mean

$$\mu_B = T \mu_A$$

and the covariance matrix

$$\Sigma_B = T * \Sigma_A * T^T$$

where T^T is the **transposed** matrix of T.

This can be applied for the LDA with the LDA matrix L and the DCT with transformation matrix C. Thus we can transform a Gaussian of the codebook (in the LDA space) to a Gaussian in the MEL frequency space (more exactly, the log MEL frequency space).

$$\Sigma_{MEL} = C^{-1} * L^{-1} * \Sigma_{LDA} * L^{-1^T} * C^{-1^T}$$

In practice, the Σ_{MEL} matrix will not be exactly the same as one that was directly estimated from the log-mel feature vectors. The reason is that the LDA and the DCT are used to reduce the dimension of the feature space. The intension of this is to get a lower dimensional feature space, that still contains the relevant information. In the rest of this section, we ignore this for simplicity. However, when implementing the technique, this aspect has to be considered. The correct treatment of dimensionality reduction is described in section 2.3.

It is clear that the mean vectors of all Gaussians remain constant in all feature spaces when only the variance of the Gaussians is changed. The Gaussian with the covariance matrix Σ_{MEL} is a Gaussian in the MEL feature space. Its dimensions correspond to frequency bands of the speech signal. Hence we can multiply this Gaussian with the G matrix as defined above, to weight different frequencies differently. This leads to

$$\overline{\Sigma_{MEL}} = G * C^{-1} * L^{-1} * \Sigma_{LDA} * L^{-1^T} * C^{-1^T} * G^T$$

Feature Space	Dimension
MEL	162
Log-MEL	162
CEP	99
LDA	32

Table 1. Example dimensions of feature space

Finally the modified Gaussian in the MEL space has again to transformed in the LDA space to be applied during the speech recognition process, leading to

$$\overline{\Sigma_{LDA}} = L * C * G * C^{-1} * L^{-1} * \Sigma_{LDA} * L^{-1^T} * C^{-1^T} * G^T * C^T * L^T$$

The used cosine transform is orthogonal, thus the $C^{-1^T} = C$, and the formula reduces to

$$\overline{\Sigma_{LDA}} = L * C * G * C^{-1} * L^{-1} * \Sigma_{LDA} * L^{-1^T} * C * G^T * C^{-1} * L^T \quad (1)$$

One final statement concerns technical aspects. The LDA in the HBAS recognizer does not only transform the feature space, but also discards some of the less relevant dimensions. In most cases, the codebook with Gaussians before the LDA is still available. In these cases, it will be better to use the codebook before the LDA modification as starting point, and perform the following operations

$$\overline{\Sigma_{LDA}} = L * C * G * C^{-1} * \Sigma_{DCT} * C * G^T * C^{-1} * L^T \quad (2)$$

Both Formula (1) and (2) can be used to generate new Gaussians optimized for the recognition of non-native speech.

2.3 Correct Treatment of Dimension Reduction

In the previous section, the fact of dimensionality reduction was ignored, thus making the assumption that all matrices are square and have the same dimension. This simplification allowed for example to write

$$\Sigma_{MEL} = C^{-1} * L^{-1} * \Sigma_{LDA} * L^{-1^T} * C^{-1^T} \quad (3)$$

without taking care about matrix dimensions. In the HBAS system however, both the DCT and the LDA are used for dimensionality reduction. Table 1 shows the size of our feature vectors after each transformation. The dimensionality reduction is achieved by discarding some dimensions when applying the DCT and the LDA. This leads to the fact that Equation (3) actually looks like Equation (4), where the matrix dimensions that are used are indicated in parentheses after the matrix. The matrix operations like transpose and invert are always executed on the square version of the matrix, however, for the multiplication only parts of the matrices have to be used.

$$\Sigma_{MEL}(162x162) = C^{-1}(162x99) * L^{-1}(99x32) * \Sigma_{LDA}(32x32) * L^{-1^T}(32x99) * C^{-1^T}(99x162) \quad (4)$$

Similarly, all Equations from Section 2.2 have to be modified for correct transformations.

2.4 MEL Bands

The previous sections have described how the influence of frequency bands can be adjusted without the need to perform a retraining. For the results, it will be interesting to relate the different bands to their actual frequency range. Most recognizers at Harman Becker are trained on 11kHz speech signals. For such a sampling rate, a 256 DFT is used on which 18 MEL bands are calculated. The frequency range of each MEL band is given above each subplot of the results in Figure 1 and Figure 2.

2.5 No Retraining needed

No matter if Formula (1) or (2) is applied, it still is arguable if the same HMMs as before can be used after the Gaussians have been changed. It is certainly true, performing a training **after** the Gaussians have been modified will lead to different HMMs. First of all, the LDA matrix itself will change, and second the alignment during the Baum-Welch algorithm will change.

Yet, we believe that not performing a retraining is the correct way, as the Gaussians, the alignment and the LDA matrix itself are all optimized for the training data. Modifying the Gaussians and performing a retraining on **native** speech will hardly give improved performance. The whole motivation why the Gaussians are modified is to account for a mismatch of the native training data to the non-native test.

3 Experimental Setup

Our semi-continuous speech recognizer uses 11 Mel Fourier Cepstral Coefficients (MFCCs) with their first and second derivatives per frame and LDA for feature space transformation. The recognizer is trained on 200 hours of US Speecon data [5]. We downsample the 16kHz data to 11 kHz. The HMMs are context dependent and semi-continuous. The codebook has 1024 Gaussians and is created with a variant of the LBG algorithm.

The native test sets consist of city names. For the non-native tests the HIWIRE data [9] is used. Our results are on the clean speech adaptation data which is provided with the data (50% of the HIWIRE data). The HIWIRE database contains English from French, Spanish, Italian and Greek speakers. The utterances are command and control in a military aeronautic scenario. The test is performed with the context free grammar provided with the data as language model.

4 Results

The performance of our algorithm depends on weights that reduce or increase the influence of the correct variances. To find good parameters we perform a grid search. Our recognizer has 18 MEL bands, and we vary the weight of each MEL band from 0.5 to 2.5 in 0.1 steps. To reduce the number of possible permutations only one weight is modified at a time, all other weights are kept at their normal value.

In our first set of plots in Figure 1 we show the effects of our modifications on the native US cities test. The plots display Word Accuracy (WA) versus the weight of the frequency band. The range of the frequency band is written above each subplot. In almost all cases, the performance is worse. This is the expected behavior, as the native training data matches the native test data.

Our second set of plots in Figure 2 presents results on non-native English by French, Spanish, Italian and Greek speakers. To facilitate a comparison, the native performance from the first set of plots is also added. The plots show a clearly different performance. In contradiction to the native case, the performance is almost always better. In many cases the results even improve over the baseline.

5 Conclusion

We have presented a technique to increase or decrease the influence of frequency bands in common HMM-Gaussian based systems. In our paper, we work on semi-continuous HMMs, but the extension to continuous HMMs should be straightforward. Once a fixed set of weights has been determined, our modification can be applied to all our recognizers within minutes, and with no retraining.

Our results show that their is some truth in our motivation, as performance differs significantly between the recognition of native and non-native speech. For non-native speech, our results show small improvements. In our limited grid search, we did not find one setting which gives best performance for all analyzed accents of English.

For future work, we believe that a more sophisticated search for a parameter setting is likely to give better overall performance. Another task left to the future is to analyze the proposed technique with the non-native accents of languages other than English. Finally, our technique could be combined with MLLR for an improved speaker adaptation. This should be helpful in the case of non-native speech.

6 References

[1] L. M. Arslan. *Foreign Accent Classification in American English*. PhD thesis, Duke University, North Carolina, 1996.

[2] N. Bodenstab and M. Fanty. Multi-pass pronunciation adaptation. In *Proc. ICASSP*, 2007.

[3] G. Bouselmi, D. Fohr, and I. Illina. Combined acoustic and pronunciation modelling for non-native speech recognition. In *Proc. ICSLP*, pages 1449–1552, 2007.

[4] R. Gruhn, K. Markov, and S. Nakamura. A statistical lexicon for non-native speech recognition. In *Proc. Interspeech*, pages 1497–1500, Jeju Island, Korea, 2004.

[5] D. Iskra, B. Grosskopf, K. Marasek, H. van den Huevel, F. Diehl, and A. Kiessling. Speecon - speech databases for consumer devices: Database specification and validation. In *Proc. LREC*, 2002.

[6] M. Raab, R. Gruhn, and E. Nöth. Non-native speech databases. In *Proc. ASRU*, pages 413–418, Kyoto, Japan, 2007.

[7] M. Raab, R. Gruhn, and E. Nöth. Codebook design for speech guided car infotainment systems. In *Proc. PIT*, pages 44–51, Kloster Irsee, Germany, 2008.

[8] M. Raab, R. Gruhn, and E. Nöth. Multilingual weighted codebooks. In *Proc. ICASSP*, Las Vegas, USA, 2008.

[9] J.C. Segura et al. The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication, 2007. `http://www.hiwire.org/`.

[10] S. Witt. *Use of Speech Recognition in Computer-Assisted Language Learning*. PhD thesis, Cambridge University Engineering Department, UK, 1999.

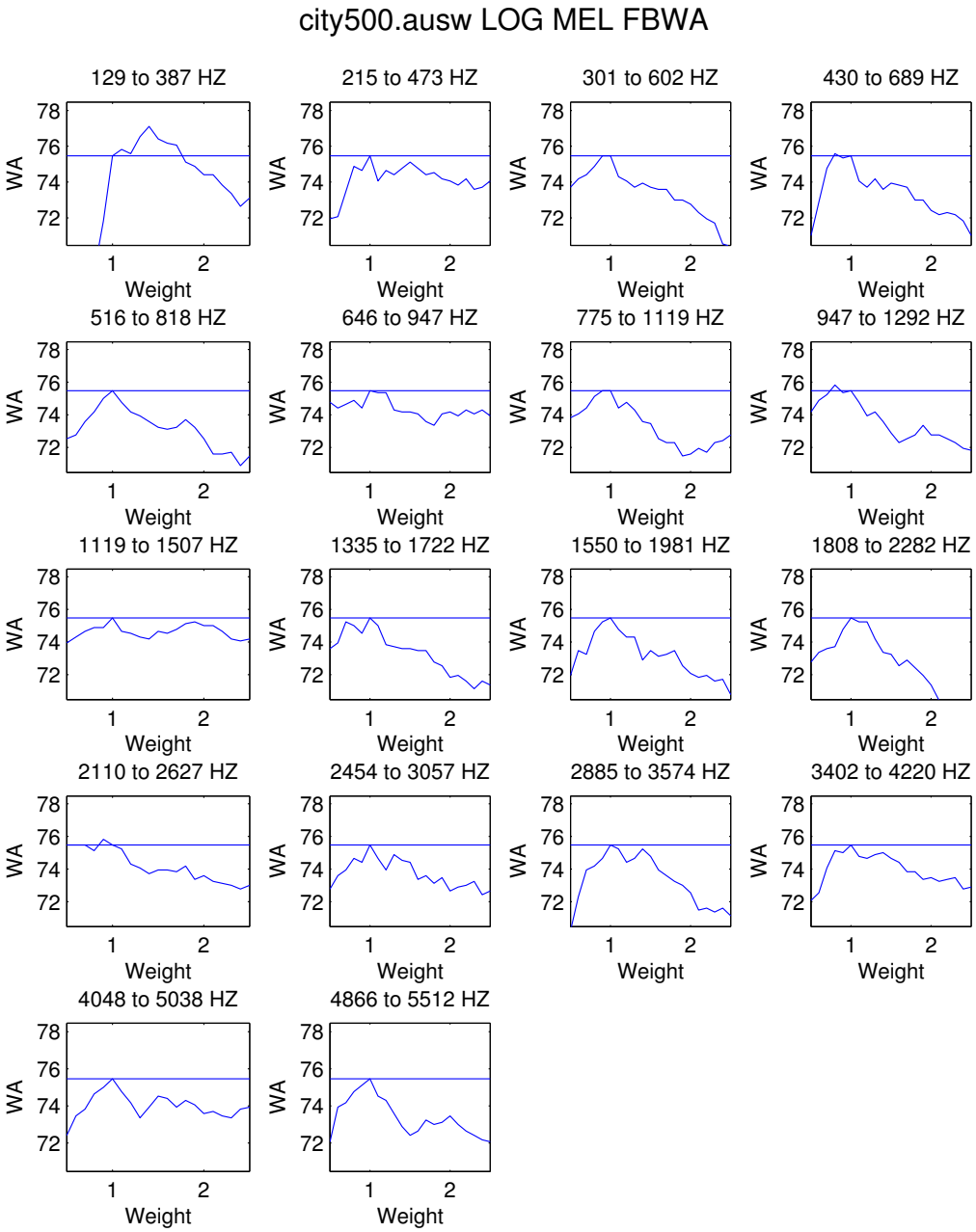

Fig. 1. FBWA on the US_City test set

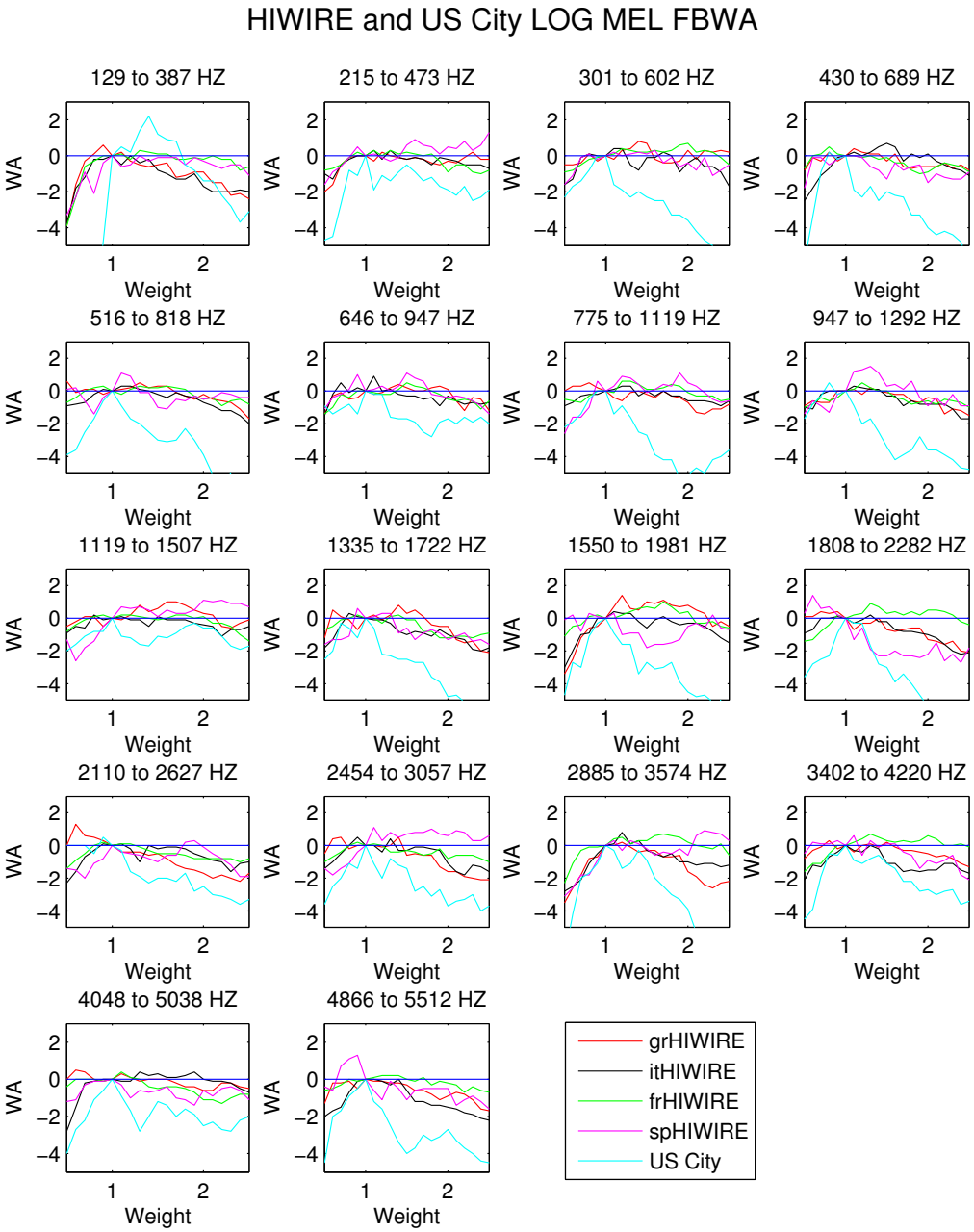

Fig. 2. FBWA on the four HIWIRE test sets