Feature Extraction for Speech Recognition

C. Lüke, K. Schnell

Institute of Applied Physics, Goethe-University Frankfurt,
Max-von-Laue-Str. 1, 60438 Frankfurt am Main, Germany
E-mail: lueke@informatik.uni-frankfurt.de, schnell@iap.uni-frankfurt.de

Introduction

One main problem of automatic speech recognition is the variability of the recorded speech signals, which is caused by the variations of the speech utterances themselves and by different environments and audio equipments. The latter is especially important for ASR-applications in natural and varying environments. Basically, three strategies can be treated to tackle this robustness problem. At first, the audio recordings themselves can be designed to capture the speech only, which can be achieved for example by microphone array techniques such as beamforming. However, this approach needs special hardware equipment. Secondly, the features for the pattern-recognition algorithms should be defined to be robust against as many kinds of distortions as possible. Finally, the speech recordings and/or their models can cover distortions from different noisy environments. The main focus of this contribution is the feature extraction. The basic features which are used here are the well-known mel-frequency cepstral coefficients (MFCCs). To improve the robustness of the feature vectors, normalization methods can be applied to the sequence of the feature vectors such as cepstral mean normalization (CMN) [1]. The attractive point of feature normalization techniques is that they are both simple and effective [2]. Especially, the simplicity makes the feature normalization techniques interesting for low-resource and small-scale applications. The normalization can balance noises and different recording influences. One assumption of the CMN is that the mean represents the environmental stationary acoustic components. However, for short utterances, the mean contains spectral characteristics of speech sounds, too. To tackle this problem, the use of a weighted mean is proposed in this contribution. The main idea of this approach is to weight the nonstationary and the stationary parts differently for the calculation of the mean. Furthermore, an optimized limiter function is introduced which is applied to the norm of the feature vectors. For evaluation, a small-scale recognition tool with a small dictionary of less than 100 words is used. For speech recognition, HMM is the state-of-the-art method for the pattern recognition problem. However, for low-resource embedded mobile applications with a very small vocabulary, the DTW approach can be sufficient, too; e.g. DTW was successful in number dialing for cell phones. Here, a DTW-based approach is used with optimized weights for the diagonal steps versus horizontal/vertical steps. In comparison to [3], different weights are used for each feature vector component.

Feature Extraction

The feature extraction describes the whole conversion from the speech signal to the sequence of feature vectors for the pattern recognition. At first, each recorded utterance is pre-processed by an automatic gain normalization using the maximum amplitude. Then, the speech signal is segmented in overlapping frames of 46ms length and 29 ms overlap, which are weighted by a Hamming window.

In the feature extraction, mel-frequency cepstral coefficients (MFCCs) are used. Firstly, the MFCCs are calculated conventionally by the DCT of mel frequency bands of the FFT. Then, the norm of the MFCC vector $\vec{x}$ is adjusted by a nonlinear function of a limiter since the distance of two feature vectors also depends on their norm. The limiter function is given by eq. (1) and includes the parameters w_g and w_L

$$\hat{\vec{x}} = \begin{cases} \vec{x}\left(\dfrac{(1-w_g)}{w_L} + \dfrac{w_g}{\|\vec{x}\|}\right) & \|\vec{x}\| < w_L \\ \dfrac{\vec{x}}{\|\vec{x}\|} & \|\vec{x}\| \geq w_L . \end{cases} \tag{1}$$

The limiter function limits the norm of each MFCC vector $\vec{x}$ that is greater than w_L, otherwise, the vector is rescaled using the parameter w_g.

The next processing step is a voice activity detection method (VAD) by simply using a threshold. Each speech part is extended by two extra frames before and after the VAD. This simple algorithm turned out to be sufficient for the evaluation of the feature extraction methods.

The complete 42-dimensional feature vectors $\vec{y}_t$ consist of:

- the logarithmic frame energy
- the delta logarithmic frame energy
- 20 MFCCs
- 20 Delta-MFCCs

Normalization methods

Now, two conventional and one specifically modified normalization method are treated. The normalization methods are applied to the whole feature vector. The cepstral mean normalization (CMN) calculates the means μ_i of the vector components $y_{t,i}$ for each utterance and subtracts them from the vector components as described by eq. (2), where i is the index of the vector component, t is the frame index representing the time, and T is the length of the utterance

$$\mu_i = \frac{1}{T}\sum_{t=0}^{T-1} y_{t,i} \qquad (2)$$
$$\hat{y}_{t,i} = y_{t,i} - \mu_i .$$

As an extension of the CMN, the cepstral variance normalization (CVN) additionally normalizes the variance by dividing each vector component by its standard variation as described by eq. (3)

$$\sigma_i^{\,2} = \frac{1}{T}\sum_{t=0}^{T-1} (y_{t,i} - \mu_i)^2 \qquad (3)$$
$$\hat{y}_{t,i} = \frac{1}{\sigma_i}(y_{t,i} - \mu_i).$$

These methods of normalization are designed for relatively long utterances in order to balance stationary noises and recording conditions. In the case of short utterances, the mean also contains spectral characteristics of individual speech sounds.

By emphasizing nonstationary regions of the utterance, we tried to tackle this problem. The norm Δy_t of the Delta-MFCC vector gives a measure of the nonstationarity of the corresponding speech frame. The nonstationarity is caused by articulatory movements of the vocal tract and by changes of the excitation.

In the following, we introduce the weighted cepstral mean normalization method (WCMN). For that purpose, the weights λ_t are defined describing the nonstationarity normalized with respect to the utterance

$$\Delta y_t = \left\| \vec{y}_t - \vec{y}_{t-1} \right\| \qquad (4)$$
$$\lambda_t = 1 + w^{\text{norm}} \cdot \frac{\Delta y_t}{\max\{\Delta y_t\}}.$$

The values of the weights λ_t are between 1 and $1 + w^{\text{norm}}$. The weights λ_t are used to calculate the weighted mean values $\tilde{\mu}_i$ as well as to scale the feature vectors. Finally, the weighted mean values are subtracted from the corresponding feature vector components

$$\tilde{\mu}_i = \frac{\sum_t y_{t,i} \cdot \lambda_t}{\sum_t \lambda_t} \qquad (5)$$
$$\hat{y}_{t,i} = y_{t,i} \cdot \lambda_t - \tilde{\mu}_i .$$

In comparison to the CVN method, the weighting factor λ_t is used for scaling. By emphasizing nonstationary frames more than stationary ones, the recognition rate can be improved in comparison to the CNM method.

DTW

For the evaluation of the feature extraction methods, the dynamic time warping algorithm (DTW) is used to compare a given utterance to reference utterances. Basically, the DTW algorithm calculates a two-dimensional distance map for pairs of utterances and determines the shortest path through the map [4].

If two samples are uttered similarly, the optimal path is not exactly the diagonal, but it is usually a slight variation of the diagonal. When trying to match utterances of different words, the shortest path calculated by the DTW algorithm is usually not diagonal. So it is reasonable to define constraints on the path in order to prefer more realistic paths. For that purpose, diagonal paths can be favored by applying a small weight w_d for diagonal steps, which penalizes concurrently horizontal and vertical steps. This can be seen from eq. (6) showing the calculation of the smallest path error to the point (t,s). $d_{t,s}$ is the distance between the feature vectors $\vec{y}_t^{\,\alpha}$ and $\vec{y}_s^{\,\beta}$ of two utterances α and β and $D_{t,s}$ is the error of the best path to the point (t,s) in the map

$$D_{t,s} = \min\begin{Bmatrix} D_{t-1,s} + d_{t,s} \\ D_{t,s-1} + d_{t,s} \\ D_{t-1,s-1} + d_{t,s} \cdot w_d \end{Bmatrix}. \qquad (6)$$

The distance map of the DTW is usually calculated by taking the standard Euclidean norm of the difference of two feature vectors. We can improve the recognition rate by applying a weight w_i^{dist} for each feature vector component

$$d_{t,s}^{\;2} = \sum_{i=1}^{42} w_i^{\text{dist}} \left(y_{t,i} - y_{s,i} \right)^2 . \qquad (7)$$

These 42 weights w_i^{dist} are equal for all utterances and can be optimized by minimizing the word error rate (WER).

In summary, depending on which method is used, there are up to 47 parameters which have to be optimized. Considering 47 parameters for the reference corpus of 439 utterances, the recognition rate can be improved.

Evaluation

The corpus consists of 439 utterances, representing a vocabulary of 58 words. These utterances are recorded in three different environments with different microphones as well as different kinds and grades of noise. Some of the utterances are whispered. These whispered utterances can be recognized only by using one of the normalization methods. In table 1, the achieved recognition rates are shown. It can be seen that the recognition rate can be improved by the WCMN method compared to the standard CMN technique. However, the CVN method still achieves better word error rates; therefore, further investigations are needed to check if a combination of the weighted normalization and the variance normalization can outperform the standard CVN.

Table 1: WER after the optimization.

WER	fixed w_i^{dist}	optimized w_i^{dist}
no normalization	9,11 %	8,66 %
CMN	8,20 %	6,38 %
WCMN	7,97 %	5,92 %
CVN	7,25 %	5,69 %

From table 1, it can be seen that the optimization of w_i^{dist} for each vector component leads to an improvement of the recognition rate. Without using the limiter, the recognition rate is significantly deteriorated. The recognition rate gets also worse if utterances of other speakers are added to the reference corpus. It is known that the DTW algorithm does not fit very well to speaker-independent ASR in comparison to the Hidden Markov Model (HMM). The figures 1-4 show an example of the utterance of the German word "Auenland"

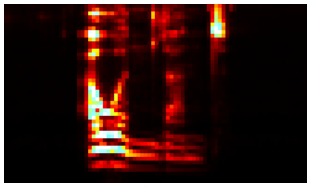

Figure 1: Mel-scaled spectrogram of the utterance "Auenland".

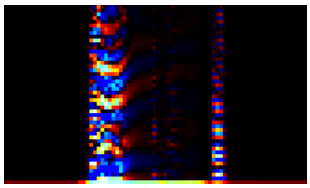

Figure 2: Cepstrogram of the utterance "Auenland".

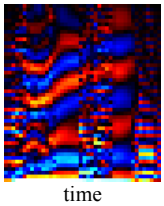

Figure 3: Cepstrogram of the utterance "Auenland" after post processing by the limiter and VAD.

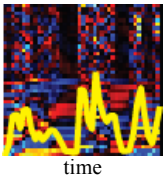

Figure 4: Feature vectors of the utterance "Auenland" with Delta-MFCCs (upper half) and normalized MFCCs (lower half); plot of the norm Δy_t of the Delta-MFCCs (yellow curve).

in spectral, cepstral, and feature-based representations over time. Positive values appear as red, negative values as blue. While in fig. 2 the central region of the utterance appears to be spectrally flat, the limiter reveals its structure, which can be seen in fig. 3. In fig. 4, the feature vectors are depicted and, additionally, the norm Δy_t of the Delta-MFCCs is plotted as a yellow curve. This curve directly correlates to the upper half of the figure, which represents the Delta-MFCCs.

Conclusion

In this paper, feature extraction for speech recognition is discussed. For that purpose, different normalization methods are evaluated using a small-size DTW-based recognition tool. To minimize the word error rate, several parameters of the feature extraction as well as the parameters of the DTW procedure are optimized. The evaluation shows that optimizing the DTW and other processing steps such as the limiter is successful regarding improvements of the recognition rate. Furthermore, the evaluation shows that the introduced WCMN achieves better results than the CMN, but is outperformed by the CVN.

References

[1] Atal, B.: 'Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification', Journal of the Acoustical Society of America, vol. 55, pp. 1304-1312, June 1974.

[2] Droppo, J.; Acero, A.: 'Environmental Robustness', in: Benesty, Jacob; Sondhi, M. M.; Huang, Yiteng (Eds.) 'Handbook of Speech Processing', Springer-Verlag, 2008.

[3] Abdulla, W.H.; Chow, D.; Sin, G.: 'Cross-words reference template for DTW based speech recognition systems', in Proc. IEEE TENCON 2003, Bangalore India, pp. 1576- 1579, 2003.

[4] Rabiner, L.; Juang, B. H.: 'Fundamentals of Speech Recognition', Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1993.