Robust Speech Recognition using Missing Feature Theory and Vector Quantization

Philippe Renevey, Rolf Vetter and Jens Kraus

Swiss Center for Electronics and Microtechnology
Neuchâtel, Switzerland
philippe.renevey@csem.ch

Abstract

This paper addresses the problem of speech recognition in noisy conditions when low complexity is required like in embedded systems. In such systems, vector quantization is generally used to reduce the complexity of the recognition systems (e.g. HMMs).

A novel approach for vector quantization based on the missing data theory is proposed. This approach allows to increase the robustness of the system against the noise perturbations with only a small increase of the computational requirements. The proposed algorithm is composed of two parts. The first part consists in dividing the spectral temporal features of the noisy signal into two subspaces: the unreliable (or missing) features and the reliable (or present) features. The second part of the proposed approach consists in defining a robust distance measure for vector quantization that compensates for the unreliable features.

The proposed approach obtains similar results in noisy conditions than a more classical approach that consists in adapting the codebook of the vector quantization to the noisy conditions using model compensation. However the computation requirements are lower in the proposed approach and it is more suitable for a low complexity speech recognition system.

1. Introduction

In this paper we are interested in increasing the performances under noisy conditions of low complexity speech recognition systems based on vector quantization. Recent works have shown that the application of missing feature theory to the problem of speech recognition in noisy conditions improves the recognition performance [1–9]. This approach consists in dividing the features representing the speech signal into a noisy speech subspace (also called missing or unreliable in the missing data theory) and a speech subspace (present or reliable) and in compensating the features according to the subspace they belong to. Several methods have been proposed [4, 9–11]. These methods require an estimation of the noise signal and involve important calculations. Therefore “classical” approaches for the detection and the compensation of the unreliable features are not suitable for low complexity systems. So we propose a novel method for both selecting present and missing subspaces and to compensate for the unreliable features. The selection of the subspaces is based on estimation of the maximum value of the noisy signal features during non-speech segments of the signal, and the compensation of the unreliable features is based on a weighting of the feature contributions during the distance measure in vector quantization. Reliable features have a high weight and unreliable features a low weight.

2. Subspace selection

In the missing feature approach, the data are divided into two subspaces, the first called missing or unreliable contains the features that are corrupted by noise, the second subspace, called present or reliable, contains features that contain mainly the speech signal. Several approaches have been proposed for the determination of these subspaces:

- **Spectral subtraction**: The spectral subtraction [12, 13] is used both for enhancement and for detection of the missing features. Frequency bands that obtain a negative energy after the subtraction are considered as missing or unreliable [4, 11].
- **Reliability measure**: A statistical model representing the noise distribution in the spectral domain is used to calculate a reliability measure of the features according to their magnitude [8].
- **Soft decision**: A similar measure for the detection of the unreliable features has been proposed in [10]. In this method, a sigmoid function is used to estimate the reliability of the features.

All these approaches require a high level of computation and specific functions like sigmoid and integration used in these algorithms are not suitable to be used in low complexity speech recognition systems.

In a low complexity system, a voice activity detector (VAD) used to detect the regions where speech is present and the recognition is performed only during the speech segments. Generally, the noise signal is estimated during the non-speech segments. In order to reduce the complexity of the algorithm, we propose to use the maximum value of the noise during the non-speech segments for the calculation of the threshold used for the selection of the reliable and unreliable features. This threshold is defined as:

$$
\Theta(\omega) = \max_{t \in \text{speech}} |Y(\omega, t)|
$$  

where \( |Y(\omega, t)| \) is the time-frequency representation of the noisy signal.

The selection of the speech and noisy speech subspaces (present/missing or reliable/unreliable) is defined as follows:

- The frequency band \( \omega \) is **reliable** if

$$
|Y(\omega, t)| > \Theta(\omega).
$$  

- The frequency band \( \omega \) is **unreliable** if

$$
|Y(\omega, t)| \leq \Theta(\omega).
$$
Frame number

Frequency band

50 100 150 200 250

5 10 15 20 25

Figure 1: Detection mask of the reliable features (black) and unreliable features (white) obtained with the reliability measure.

Figure 2: Features representing the sequence of digits “2409829” in babble noise with a SNR of 6 dB.

Figs. 1 and 3 present the division between reliable and unreliable features for the sequence of digits “2409829” in babble noise with a SNR of 6 dB (Fig. 2).

The results obtained by the two methods of detection are similar, but the proposed method offers the advantage of the lowest computation requirements and that the estimation of the maximum of the noise is more easy and more accurate than the estimation of its distribution.

3. Vector quantization with unreliable features

In vector quantization, the problem is to find the best centroid to represent the feature vectors. They are generally selected using a distance measure, the centroid with the smallest distance is selected and the feature vector is simply represented by the index of the centroid. In this paper we used the Mahalanobis distance:

$$D^2_{M}(x, i) = (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)$$

(4)

where $\mu_i$ is the centroid $i$ and $\Sigma_i$ is the covariance matrix.

In this paper we chose to use a diagonal covariance matrix in order to reduce the complexity of the system. So the Mahalanobis distance is expressed as:

$$D^2_{M}(x, i) = \sum_{\omega=1}^{\Omega} \frac{(x(\omega) - \mu_i(\omega))^2}{\sigma^2_i(\omega)}.$$  

(5)

The problem is to find a robust approach to compute this distance in order to provide a noise robustness for the vector quantization, without an important increase of the complexity of the recognition system. Therefore we propose to adapt the methods developed in the framework of the missing data theory in order to increase the robustness of the vector quantization. Classical missing feature approaches generally process the unreliable features using one of the two possible solutions. The first consists in the integration of the probabilities of the HMMs over the possible values that the unreliable features should have. Two kinds of integration methods have been presented, the integration over the whole interval (from $-\infty$ to $\infty$) and the integration over a restricted interval. These techniques are generally called marginalisation and bounded integration. The second solution for the compensation the unreliable features, called imputation consists in replacing the unreliable features by their most probable values using the information of the reliable features and some statistical model.

These two solutions require memory and computation power that is not suitable for the design of a low complexity speech recognition system. Therefore we introduce a new method for the compensation of the unreliable features.

This method is based on three assumptions:

1. **Speech and noise are additive in the magnitude spectral domain.**
2. **The highest the magnitude of the noisy signal is, the smallest is the relative influence of the noise.**
3. **Under the assumption of additivity, features with a small magnitude are more representative of the speech signal than of the noise signal.**

All these assumptions are approximative, but are often used in enhancement methods like spectral subtraction for example. From these assumptions we derive two rules for the compensation of the unreliable features:

...
To compare the performance of the proposed approach, we used an algorithm that is based on the compensation of the vectors representing the centroids of vector quantization. In this algorithm, the noise is considered as normally distributed in each frequency band. These distributions are estimated during the non-speech segments. The codebook used for the vector quantization is compensated using a log-normal approximation like in parallel model combination (PMC).

Table 3: Recognition results for factory noise

<table>
<thead>
<tr>
<th>SNR dB</th>
<th>B % acc</th>
<th>SW % acc</th>
<th>CC % acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>35.6</td>
<td>60.6</td>
<td>57.8</td>
</tr>
<tr>
<td>12</td>
<td>49.4</td>
<td>76.7</td>
<td>78.1</td>
</tr>
<tr>
<td>18</td>
<td>60.2</td>
<td>85.9</td>
<td>86.4</td>
</tr>
<tr>
<td>24</td>
<td>64.9</td>
<td>88.5</td>
<td>88.7</td>
</tr>
</tbody>
</table>

Table 4: Recognition results for F16 noise

<table>
<thead>
<tr>
<th>SNR dB</th>
<th>B % acc</th>
<th>SW % acc</th>
<th>CC % acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>9.6</td>
<td>36.9</td>
<td>34.1</td>
</tr>
<tr>
<td>12</td>
<td>30.3</td>
<td>63.1</td>
<td>61.71</td>
</tr>
<tr>
<td>18</td>
<td>47.4</td>
<td>77.5</td>
<td>77.9</td>
</tr>
<tr>
<td>24</td>
<td>56.6</td>
<td>86.5</td>
<td>85.9</td>
</tr>
</tbody>
</table>

5. Results

Tables 1-4 shows the recognition scores obtain with four noises (babble, car, factory and F16 noises) for SNRs ranging from 6 to 24 dB. The three algorithms are:

- **B**: Baseline, the original vector quantization.
- **SW**: Score Weighting, the proposed algorithm.
- **CC**: Codebook Compensation, the compensation of the codebook using a noise model.

We observe that the proposed algorithm and the codebook compensation algorithm obtain similar performances in all the noise conditions and offer a significant improvement when compared with the baseline system. The worst results are obtained in babble noise, explained by the fact that the noise has the same spectral nature than the speech signal and therefore the selection and the compensation of the speech and noisy speech subspaces is more difficult. On the other side, the results obtained in car noise environment are very good. It is also explained by the spectral nature of the noise. In this case, the energy of the noise is concentrated in low frequencies and is easy to compensate.

6. Conclusion

In this paper we have proposed a novel approach for robust speech recognition in the framework of missing features and vector quantization. The aim of this approach is to design a robust speech recognition algorithm for low complexity speech recognition system. We proposed an algorithm for the determination of the noisy speech and speech subspaces based on an estimation of the maximum value of the noise in each subband. This algorithm offers the advantage to be far less complex than previously proposed algorithms and the performance.
of the selection can be compared with the classical approaches. We also proposed a new method for the compensation of the unreliable features based on a weighted sum of the contribution of each sub-band in the calculation of the distance during the vector quantization. Finally we showed that the performance of the proposed approach are similar to those of a more complex approach. Therefore the proposed approach is well suited for increasing the performance of vector quantization based speech recognition systems in adverse conditions.

7. References


