A New Data Driven Method for Robust Speech Recognition

NASER AYAT MOHSEN RAHMANI
Computer Group
Payame Noor University
Arak University

Abstract: The conventional view on the problem of robustness in speech recognition is that performance degradation in ASR systems is due to mismatch between training and test conditions. If problem of robustness in ASR systems were considered as a mismatch between the training and testing conditions the solution would be to find a way to reduce it. Common approaches are: Data-Driven methods such as speech signal enhancement and using robust features and model-based methods that alliterate or adapt model of speech signal. In this paper, we review some of data-driven and model-based methods and implement some of data-driven methods such as: spectral subtraction, cepstral mean normalization, cepstral mean and variance normalization, mapping noisy space to clean space and SNR-dependent cepstral normalization. We suggest a new data-driven method for robust speech recognition using neural networks and compare it to other methods.

Key-Words: Robust speech recognition, cepstral normalization, neural networks

1 Introduction
The conventional view on the problem of robustness in speech recognition is that performance degradation in ASR systems is due to the differences between speech signal they receive on input (when employed in real life applications) and the speech signal used for training and estimation of parameters of their models during system construction. This is commonly referred to as mismatch between training and test conditions. Some common reasons for mismatch between training and testing speech signal, are considered to be: contamination of signal with noise (additive, convolutional, reverberation), speaking style (Lombard effect, speaking rate) and inter speaker variations (voice quality, pitch, gender). If problem of robustness in ASR systems against contamination with noise, were considered as a mismatch between the training and testing conditions the solution would be to find a way to reduce it. Common approaches are: speech signal enhancement, using robust features, using microphone arrays, using hearing properties of human ear and model alteration/adaptation.

The suggested techniques for description of effects of noisy environment can be divided in two categories: data-driven methods and model-based methods. Data-driven methods try to describe effects of environment on speech and speech features and enhance speech signal or improves its features. Some of these methods need to both of noisy and clean signal. Model-based methods try to change statistical model of environment such as it adapts to new properties of environment. In these methods, the observation is not changed and there is not any assumption or change for speech signal. Model-based methods modify acoustic models instead of speech signal or its features. This has the advantage that no decisions or hypotheses about speech are necessary and observed data is unaltered. These methods don’t need to stereo database. Some examples of these approaches are: Hidden Markov Model (HMM) decomposition[7,14,15], Parallel model combination (PMC)[2,3] and maximum likelihood regression (MLLR)[2,3,8].

In this paper, we limit ourselves to data driven-methods. After reviewing basic methods, we evaluate performance of them and reap profits of reviewed
methods. In second section, a model for speech recording at the input of an ASR device is discussed. This model considers acoustic noise and transmission channel effects and can explain the relation among different approaches of robust ASR. Third section is dedicated to discussing of data-driven methods. Fourth section includes our proposed method. Our experiments and results are explained in section five. In section six, we give our conclusion.

2 A model of Environment

In most of speech recognition applications, there are two kinds of noise: additive noise and linear filter noise. Additive noises include background sounds, effects of air flow and unwanted signals captured by microphone. Linear filter noises include effects of microphone or transmission channel and reverberation. If kinds of noise are limited to additive and channel noise, we can construct a model such as figure (1), for effects of environment [1,3].

According to [3], after transforming to log spectral domain, relation (1) is converted to:

\[ y = x + h + \ln(1 + e^{n-x-h}) \]

\[ = x + h + s(x, n, h) \]  (2)

Where \( y, h, n, x \) represent the log spectrum of \( y[m], h[m], n[m] \) and \( x[m] \) respectively. Equation 2 can be rewritten as:

\[ y = x + f(x, n, h) \]  (3)

Where \( f \) can be written as:

\[ f(x, n, h) = h + \ln(1 + e^{n-x-h}) \]  (4)

Equation (4) represents feature vectors in log-spectral domain. As the relationship between cepstral vectors and log-spectral vectors is linear, we can write:

\[ c_y = c_x + g(c_x, c_n, c_h) \]  (5)

Where \( c_y, c_x, c_n, c_h \) are cepstrum of \( y[m], x[m], n[m] \) and \( h[m] \) respectively. Function \( g \) can be computed as the inverse Fourier transform of function \( f(x, n, h) \):

\[ g(c_x, c_n, c_h) = c_h + \left[ \ln(1 + e^{c_x - c_n - c_h}) \right] \]  (6)

In this relation, \( g \) is an additive term to the environment that depends to \( c_x, c_n, c_h \). This is a complex function and not so easy to compute. In effect, the noise and channel features are unknown. In following sections, we review different approaches proposed for estimating clean feature \( c_x \) from noisy features \( c_y \). Finally we propose a method for learning \( g(.) \) using neural networks. The proposed method simply learns \( g(.) \) and maps noisy features to clean features.

3 Backgrounds

Data driven methods determine effects of environment on properties of speech signal and try to improve these effects such as speech is recognized in an environment independent manner. These methods can be divided to multiple groups. First group acts on noisy speech signal directly and estimates clean signal from noisy signal. Spectral subtraction and signal filtering are placed in this group [17].
Second group is based on finding and extracting robust features. Methods of this group try to remove effects of noise in features and reduce mismatch of training and testing data. CMS\(^1\), PLP\(^2\) and RASTA PLP \[4,5\] and ZCPA\(^3\) \[6\] analysis are examples of these methods. Another group finds a map between noisy signal features and clean signal features. For estimating such mapping, we must access to both of clean and noisy signals. There is much kind of maps: Linear regression methods \[18\], nonlinear estimators such as MLP or other neural networks \[18\] and minimizing a cost function by MMSE criterion. In fourth approach, methods of pervious groups are combined. Properties of human ear are also used to improve robustness of ASR methods. Noise masking technique is an example of such methods \[2,7\].

### 3.1 Spectral Subtraction

Spectral subtraction is a well-known method that has a long in research in speech enhancement. In this technique, an estimate of the noise spectrum \( (P_n(\omega)) \) is computed and it is subtracted from noisy input spectrum \( (P_x(\omega)) \) to obtain an estimation of clean speech spectrum.

\[
\hat{P}_x(\omega) = P_x(\omega) - P_n(\omega) \tag{7}
\]

Or

\[
\hat{P}_x(\omega) = \begin{cases} 
  P_x(\omega) - \alpha \cdot \hat{P}_n(\omega) & P_x(\omega) > \beta \hat{P}_n(\omega) \\
  \beta \cdot \hat{P}_x(\omega) & \text{otherwise}
\end{cases} \tag{8}
\]

Where \(\alpha\) and \(\beta\) are overestimation and noise floor parameters.

Spectral subtraction can be used as a preprocessing step for ASR system to make them robust against quasi time-invariant noise.

Many different methods are developed using of spectral subtraction. Adaptive spectral subtraction \[16\] and nonlinear spectral subtraction \[16\] are examples of such methods.

### 3.2 Cepstral Mean Subtraction

Cepstral Mean Subtraction (CMS) is perhaps one of most effective algorithms considering its simplicity. It is applied in most large vocabulary speech recognition systems. This algorithm computes a long-term mean value of feature vectors and subtracts this mean value from each of the vectors. This helps in reducing variability of data and does a kind of normalization and this is reason of naming it as Cepstral Mean Normalization. This procedure is applied to both of training and testing data. If purely convolutional noise is present, this set of speech parameters is unaffected by changes in noise. When additive noise is present, subtracting the mean is found to aid the robustness of system. However, the system behavior is hard to predict when both forms of noise are present \[2,3\].

### 3.3 SNR-Dependent Cepstral Normalization (SDCN)

SNR-dependent cepstral normalization is an algorithm that operates directly in cepstral domain by adding a compensation vector that depends exclusively on SNR of input frame. If \(v, x, y\) represent, compensation vector, cepstral vector of noisy signal and cepstral vector of clean speech respectively, we can write:

\[
y = x + v \tag{9}
\]

We determine compensation vector as a function of SNR and then we can write:

\[
y = x + v(SNR) \tag{10}
\]

Inspection of equation (2), indicates that at high SNR, \(x + h \geq n\) and \(s \approx 0\) and then \(y \approx x + h\) and at low SNR, \(x + h \leq n\) and \(y \approx n\) Hence, the SDCN algorithm performs spectral equalization at high SNR and noise suppression in low SNR and at intermediate SNR, it can be an approximation \[1,9,10\]. In estimation of compensation vector \(v \text{ (SNR)}\), the goal is to transform features of noisy speech signal so that it looks like the features of clean speech signal. According to this criterion, the correction vectors are estimated by computing the average

---

1. Cepstral Mean Subtraction
2. Perceptual Linear Predictive analysis of speech
3. Zero Crossing with Peak Amplitude
difference between cepstral vectors for noisy speech signal versus cepstral vectors for clean speech signal on a frame-by-frame basis as a function of input SNR. Equation of estimation is as follows [1]:

$$\sum_{i=0}^{N}\sum_{j=0}^{1}\delta\{SNR_i - k\Delta_{SNR}\} = \sum_{i=0}^{N-1}\delta\{SNR_i - k\Delta_{SNR}\}$$

Where $x_i[j]$, $y_i[j]$ represent element $j$ of cepstrum vectors at frame $i$ for the clean and noisy speech signal respectively. $SNR_i$ is the SNR of frame $i$ in noisy speech signal. $\delta$ is Kronecker delta and sum is carried out for the entire $N$ frames in database.

SDCN algorithm requires a stereo database of our standard environment and the new environment to train the correction vectors and then SDCN algorithm is not environment-independent.

### 3.4 Mapping from noisy space to clean space

In mapping approach, the goal is finding a function or a transformation to transforms noisy feature to clean speech signal feature approximately. There are different mathematical methods for this goal and one of most important of them is neural network.

#### 4 Proposed method

In this paper, we use mapping in cepstral domain and clean cepstral coefficients are estimated from noisy cepstral coefficients by a neural network. Our neural network is a MLP with 4 layers. Each of two hidden layers has 20 nodes and each of input and output layers have 12 nodes. Figure 4 shows samples of clean, noisy and estimated cepstral features by MLP.

![Architecture of MLP for estimation of clean cepstral coefficients](image)

In this work, we use neural networks for mapping of robust features from noisy space to clean space. This approach is a combination of two methods: mapping and robust features and it needs Stereo database.

![Block diagram of mapping in transform domain](image)

Mapping can be done in different domains. Mapping in transform domain maps noisy signal to clean signal. [13] With respect to nonlinear adding of noise and speech in new domains, isolation must be done in a nonlinear approach. Neural networks can be used as a nonlinear isolator. For different mappings in different SNRs, an estimated value of noise or SNR is used as one of neural network input. Figure 2 shows a block diagram of mapping in transform domain.
Fig. 4: (a) first cepstral coefficients for 25 frames (b) 12 cepstral coefficients in one frame (solid: clean feature - dashed: noisy feature – dotted: estimated feature by MLP)

5. Experiments and Results
In this section, we explain details of implementation for some of techniques that we have described in third section and we compare them to our proposed technique. Our recognition system is continuous density hidden markov model (CDHMM), with 6 states and 2 Gaussian mixtures in each state. Our database is Persian numbers 1 to 10. A male speaker has spoken them and there is 15 utterance of each number. Each number has been recorded by sampling rate 16 kHz and 16 bits per sample, in a clean environment. Ten utterances are used as training data and 5 utterances is used as testing data. Our feature vector contains 12 MFCC coefficients and their first order derivative and logarithm of energy and its first order derivative. Hence, length of feature vector is 26. We have used 30 ms frames and 15 ms overlap and Hanning window. We have used additive white Gaussian noise. This noise is selected from NOISEX92 database and downsampled to 16 kHz .We have used different SNRs: 0 db, 5 db, 10 db, 20 db, 30 db. Our Recognition rate for clean testing data was 98.7%. The implemented methods and their results are shown in table1.

<table>
<thead>
<tr>
<th>Method</th>
<th>SNR 30 db</th>
<th>SNR 20 db</th>
<th>SNR 10 db</th>
<th>SNR 5 db</th>
<th>SNR 0 db</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>96.7%</td>
<td>74.7%</td>
<td>24%</td>
<td>15.3%</td>
<td>13.3%</td>
</tr>
<tr>
<td>MFCC+NN</td>
<td>98%</td>
<td>92.7%</td>
<td>69.3%</td>
<td>36.6%</td>
<td>25.3%</td>
</tr>
<tr>
<td>CMS</td>
<td>96.7%</td>
<td>93.3%</td>
<td>74.6%</td>
<td>47.3%</td>
<td>12%</td>
</tr>
<tr>
<td>CMVS</td>
<td>96.7%</td>
<td>85.4%</td>
<td>40%</td>
<td>27.5%</td>
<td>22.7%</td>
</tr>
<tr>
<td>CMS+NN</td>
<td>97.3%</td>
<td>95.3%</td>
<td>80.1%</td>
<td>52%</td>
<td>36%</td>
</tr>
<tr>
<td>SS</td>
<td>98.4%</td>
<td>96.2%</td>
<td>64.3%</td>
<td>37%</td>
<td>21.7%</td>
</tr>
<tr>
<td>SDCN</td>
<td>96.8%</td>
<td>89.9%</td>
<td>69.1%</td>
<td>35.3%</td>
<td>25%</td>
</tr>
</tbody>
</table>

Table 1-recognition result for different data-driven robust method

Estimating of clean CMS by MLP has best performance and performance of CMVS is acceptable.

Short length of utterance and insufficient number of frames can cause the unacceptable performance of CMVS in our experiments. Increasing of distortion and musical noise in low SNR (especially 0 db) decreases performance of spectral subtraction. CMS is more effective for convolutive noise and it can cause low performance of CMS in SNR= 0 db.

As mentioned before, SDCN compensates effect of convolutive noise on cepstral features in high SNR and effect of additive noise on cepstral features in low SNR. It has similar behavior in our experiments and in SNR=0 db, its performance is better than most of other methods.

MFCC, CMS, CMVS and SS methods don’t need to stereo data and other methods need to both of clean and noisy signal and this is a limitation for their application in real conditions.

CMVS: cepstral normalization using mean and variance
CMS+NN: estimating CMS of clean signal using MLP
SS: spectral subtraction with overestimation 1.5 and noise flooring parameter 0.3
SDCN: SNR-dependent cepstral normalization.

In SNR = 30 db, estimating of clean Mel cepstrum by MLP and spectral subtraction have best performance. In SNR=20 db, 10db, 5 db, cepstral mean subtraction and estimating of it by MLP have best performance, where noise is additive .In SNR=0db,
6. Conclusion
Compensation methods for effects of noise in environment can be divided to two categories: data-driven methods and model-based methods. Data-driven methods modify speech signal or its parameter and model-based methods modify acoustic models instead of speech signal or its features. A part of data-driven methods need to stereo database and this can be limited their practical application. Model-based approaches have not such limitation. Some of data-driven methods is discussed and implemented in our work and a new approach is presented. In our approach, robust feature and mapping group is combined and we use MLP for mapping noisy cepstral features to clean cepstral features and result is an estimated value for clean cepstral features. This approach has best performance between all other methods, which we have implemented. An Important note is that result is in case of additive noise. This approach needs to stereo database and then, future works must focus on methods or modified methods that have not such limitation.

References