Evaluation of Universal Compensation on Aurora 2 and 3 and Beyond

Ji Ming†, and Baochun Hou†

†School of Computer Science, Queen’s University Belfast, Belfast BT7 1NN, UK
‡School of Electronic, Communication and Electrical Engineering
University of Hertfordshire, Hatfield AL10 9AB, UK

Abstract

A new method, namely universal compensation (UC), is introduced for speech recognition involving additive noise assuming no knowledge about the noise. The UC method involves a novel combination of the principle of multi-condition training and the principle of the missing-feature method. Multi-condition training is employed to convert full-band spectral corruption into partial-band spectral corruption through compensations for simulated wide-band noise, and the missing-feature principle is employed to reduce the effect of the remaining partial-band corruption on recognition by basing the recognition mainly on the matched or appropriately compensated spectral components. This combination makes the new method potentially capable of dealing with any additive noise – with arbitrary temporal-spectral characteristics – based only on clean speech training data and simulated noise data, without requiring knowledge about the noise. This paper describes the evaluation of the UC method on Aurora 2 and 3 and further, on noise conditions unseen in the Aurora tasks. The results show that the new model assuming no knowledge of noise has performed equally well as the baseline models trained for the specific tasks. The new model has outperformed the baseline when there exists a mismatch between the training and testing conditions.

1. Introduction

It is important that speech recognition systems can maintain high accuracy in the presence of acoustic corruptions (noise) such as those introduced by the speaker’s environment and by the channel through which the signal propagates. To date much research has targeted the impact of noise through filtering techniques such as spectral subtraction or Wiener filtering [1], assuming a priori knowledge of the noise spectrum. Other techniques rely on robust features, such as those produced by cepstral mean removal or RASTA [2], which provide robustness to channel distortion and/or slowly-varying noise. Speech recognition systems incorporating a statistical model for the noise or training data from the noisy environment have been developed, for example, the system based on PMC (parallel model combination) [3] and the systems trained for specific tasks such as in-car speech [4]. Recent research in SPLICE has shown improved noise robustness by modeling corruption characteristics based on stereo training data [5]. Recent studies on the missing-feature method has assumed that noise only causes a partial corruption to the speech representation and that the remaining part of the representation not affected by the noise can be exploitable [6]–[8].

This paper describes a new method as a complement to the above research. We study speech recognition involving additive background noise, assuming any corruption type (e.g., full, partial, stationary or time-varying), and assuming no knowledge about the noise characteristics and no training data from the noisy environment. The new method involves a novel combination of the principle of multi-condition training and the principle of the missing-feature method. Multi-condition training is employed to convert full-band spectral corruption into partial-band spectral corruption through compensations for simulated wide-band flat-spectrum noise at consecutive signal-to-noise ratios (SNRs), and the missing-feature principle is employed to reduce the effect of the remaining partial-band corruption on recognition by basing the recognition mainly on the matched or appropriately compensated spectral components. The combination of these two strategies makes the new method potentially capable of dealing with any additive noise – with arbitrary temporal-spectral characteristics – based only on clean speech training data and simulated noise data, without requiring knowledge about the noise. For this reason, we term the new method universal compensation (UC). A previous study of the UC technique was described in [9]. In this paper, this study is deepened by a further theoretical development, and by a full evaluation of the model on the Aurora 2 and 3 tasks and beyond, on noise conditions unseen in Aurora.

2. Universal Compensation

The UC technique includes three steps:

1. Construct a set of models for short-time speech spectra using artificial multi-condition speech data, consisting of the clean training data and a collection of noisy training data generated by corrupting the clean training data with artificial wide-band flat-spectrum noise at consecutive SNRs;

2. Given a test spectrum, search for the spectral components in each model spectrum that best match the corresponding spectral components in the test spectrum, and produce a score based on the matched components for each model spectrum;

3. Combine the scores from the individual model spectra to form an overall score for recognition.

These three steps may be explained using a simple example, shown in Fig. 1, assuming a single short-time spectrum (i.e. a frame). Fig. 1 shows, on the left-hand side, an instance of a clean speech spectrum, representing the data available for training. Wide-band flat-spectrum noises with different SNRs are added, respectively, to the waveform of the clean frame, to form

This work was supported by the UK EPSRC grant GR/S63236.
the set of noisy training data, i.e. Step 1. The noise may be generated by passing a white noise through a low-pass filter with the same bandwidth as the speech spectrum. Assume that this leads to a set of model spectra, shown in the middle of Fig. 1, each model spectrum corresponding to a specific SNR, and including an appropriate compensation for a wide-band flat-spectrum noise at that SNR. The clean spectrum is also included in the model set (shown at the top of the model spectra). Fig. 1 shows, on the right-hand side, an example of a test spectrum, which is assumed to be the result of the clean frame with the addition of some noise. The characteristic of the noise spectrum can be arbitrary and is not known a priori. While the test spectrum involves a full-band corruption with respect to the clean spectrum, it involves only a partial-band corruption when compared to some of the model spectra, for example, (from top to bottom), model spectra 2, 3, 4 and 5 in Fig. 1, assuming that a local frequency-band distortion in the test spectrum due to the addition of a noise may be matched by the corresponding model spectrum with the addition of a “flat-spectrum” noise in the same frequency band with a similar SNR. These matched parts, for this particular example, are enclosed within the circles over the appropriate model spectra as shown in Fig. 1. Thus, the step of comparing the test spectrum with each model spectrum to find their matched components effectively results in a conversion of a full-band corruption to a series of partial-band corruptions, assuming that the test spectrum involves only a partial-band corruption when compared to at least one of the model spectra. The effect of partial-band corruption on recognition can be reduced by ignoring the distorted spectral components. This is achieved in Step 2 by calculating a score for each model spectrum based only on the matched spectral components. Finally, the scores from the individual model spectra are combined to produce an overall score, indicating the probability of the test spectrum associated with the model, i.e. Step 3. Note that a partial-band corruption remains partial in this conversion.

Use of artificially added noise at various SNRs to account for unknown noise sources is not new and has been discussed previously by a number of researchers (e.g. [10]). The UC method is new in that it combines artificial noise compensation with the missing-feature method, to accommodate mismatches between the simulated noise condition and the actual noise condition. This combination makes it possible to accommodate sophisticated spectral distortion, full, partial, white, colored or none, with simulated noises of a limited variety, for example, the wide-band flat-spectrum noises with different SNRs.

3. Acoustic Modeling

Formulating the UC method is straightforward following the above example. Assume that L levels of SNR are chosen to generate the wide-band flat-spectrum noises to form the noisy training data, and that each model spectrum is modeled by a probability distribution for its spectral components. Denote by $p(x|s, l)$ the probability distribution for a model spectrum associated with speech state $s$ and trained for SNR level $l$, $l = 1, 2, ... , L$. For convenience, we address each model spectrum by its index $(s, l)$.

Assume that each short-time spectrum or frame consists of $N$ spectral components. Let $o = (o_1, o_2, ..., o_N)$ be a test spectrum, which may be corrupted by noise but knowledge about the noise spectrum is not available. Recognition involves classifying each test spectrum $o$ into an appropriate speech state $s$, based on the matching components between the test spectrum and each of the model spectra $(s, l)$ associated with state $s$. Denote by $o(s, l)$ the subset in the test spectrum $o$ containing all the matching components for model spectrum $(s, l)$. The subset $o(s, l)$ is a function of the model spectrum $(s, l)$ being compared, with both the size and components being variable from model spectrum to model spectrum. Given $o(s, l)$ for each model spectrum $(s, l)$, the overall probability of $o$, associated with speech state $s$, can be defined as

$$p(o|s) = \sum_{l=1}^{L} w(s, l)p(o(s, l)|s, l) \tag{1}$$

where $p(o(s, l)|s, l)$ is the probability of $o(s, l)$ associated with model spectrum $(s, l)$, representing an estimate of the probability of $o$ associated with model spectrum $(s, l)$ with the most severely mismatched components ignored, and $w(s, l)$ is a weight, for the contribution from the corresponding model spectrum. As described in Step 2, the probability for a model spectrum $(s, l)$ is calculated based on the matched components between the model spectrum and the test spectrum. For simplicity, we assume that the individual spectral components are independent of one another. So the probability $p(o_{sub}|s, l)$ for any subset $o_{sub} \in o$ can be written as

$$p(o_{sub}|s, l) = \prod_{o_n \in o_{sub}} p(o_n|s, l) \tag{2}$$

where $p(x_n|s, l)$ is the probability distribution of the $n$th spectral component associated with model spectrum $(s, l)$.

Equation (1) is reduced to the standard mixture model when all spectral components from the test spectrum are involved in the computation (i.e., $o(s, l) = o$). This mixture model involving all spectral components is used for the training data, to model speech spectra without missing components. This model is estimated on the training set consisting of both clean data and the artificial noisy data involving wide-band flat-spectrum corruption with different SNRs. This estimation can be carried out in the same way as a conventional mixture model using the standard EM algorithm.

Given the model, computing the mixture probability in (1) using only a subset of data for each of the mixture densities is required in testing for reducing the effect of uncompensated noisy spectral components on recognition. To achieve this, we need to decide, for each model spectrum $(s, l)$, the subset $o(s, l) \in o$ that contains all the matching components. In principle, the traditional missing-feature methods concerning the identification of corrupt data, based on an estimate of the local data reliability, can be used to tackle this problem. In this
paper, we consider a solution to the problem by maximizing the appropriate probabilities. If we can assume that the matched subset produces a large probability, then \( p(s, l) \) may be defined as the subset \( o_{sub} \) that maximizes the probability \( p(o_{sub}|s, l) \) among all possible subsets in \( o \). However, (2) indicates that the value of \( p(o_{sub}|s, l) \) is a function of the size of the subset \( o_{sub} \). Implied is that the values of \( p(o_{sub}|s, l) \) for different sized subsets are of a different order of magnitude and are thus not directly comparable. An appropriate normalization is needed for the probability \( p(o_{sub}|s, l) \), to make the contributions from different sized subsets comparable, and to favor the subset with the largest number of matching components since this provides maximum information for discrimination. A possible solution to this is to replace the conditional probability of the test subset, \( p(o_{sub}|s, l) \), with the posterior probability of the model spectrum, \( p(s, l|o_{sub}) \). The posterior probability of model spectrum \( (s, l) \) given a test subset \( o_{sub} \), is defined as

\[
p(s, l|o_{sub}) = \frac{p(o_{sub}|s, l)p(s, l)}{\sum_{s', l'} p(o_{sub}|s', l')p(s', l')} \tag{3}
\]

where \( p(o_{sub}|s, l) \) is the conditional probability of test subset \( o_{sub} \). Given model spectrum \( (s, l) \), as defined in (2), and \( p(s, l) \) is the prior probability for model spectrum \( (s, l) \). The posterior probability \( p(s, l|o_{sub}) \) defined in (3) is normalized for the size of the test subset, always producing a value in the range \([0, 1]\) for any sized \( o_{sub} \). Most importantly, it can be shown that this posterior probability favors large matched subsets, i.e., it produces larger values for the subsets containing larger numbers of matched components. Thus, by maximizing the posterior probability \( p(s, l|o_{sub}) \) with respect to \( o_{sub} \), we should be able to obtain the subset for model spectrum \((s, l)\) that contains all the matched components in terms of the maximum a posteriori (MAP) criterion. The following shows the optimum decision:

\[
o(s, l) = \arg \max_{o_{sub}|s} p(s, l|o_{sub}) \tag{4}
\]

Since the conditional probability \( p(o_{sub}|s, l) \) and posterior probability \( p(s, l|o_{sub}) \) are proportional to each other, we replace \( p(o(s, l)|s, l) \) in (1) by the optimized posterior probability in (4), obtaining a modified version of (1) used for recognition:

\[
p(o|s) \propto \sum_{l=1}^{L} w(s, l) \max_{o_{sub}|s} p(s, l|o_{sub}) \tag{5}
\]

Equation (5) can be incorporated into a hidden Markov model (HMM), by using \( p(o|s) \) as the state-based emission probability for frame vector \( o \) associated with state \( s \).

4. Experimental Evaluation

This section describes the evaluation of the new UC model, outlined above, firstly on the Aurora 2 and Aurora 3 databases and further on noise conditions unseen in the Aurora tasks. In all the experiments, the UC model was trained using only the “clean” speech data provided within the training set of each database (details are given below for each specific database). The clean training set is expanded by adding wide-band flat-spectrum noise to each of the training utterances at ten different SNR levels, starting with SNR = 20 dB, reducing 2 dB every level, until SNR = 2 dB. The model may be estimated by explicitly associating the training spectral vectors of the same SNR with a specific probability density, as implied in (1). Alternatively, the model may also be estimated by pooling the training vectors of all SNRs together, and letting the EM algorithm decide the association between the data and the probability densities automatically. The second method was used in the experiments. Specifically, each word (digit) was modeled by a 15-state HMM, with 32 diagonal Gaussian mixture components in each state (i.e., \( L = 32 \) in Equation (1)) to account for the expanded training set including both the clean data and the artificial noisy data with ten SNR levels. The speech was divided into frames of 25ms at a frame rate of 10ms. Each frame was featured using 12 log filter-bank amplitudes, decorrelated by a high-pass filter \( H(z) = 1 - z^{-1} \). The first-order and second-order delta parameters were appended, thus forming a 36-element feature vector for each frame.

4.1. Aurora 2

Aurora 2 consists of English digit sequences with artificially introduced additive noise and simulated channel distortion. Eight different noises are included: subway, bubble, car, exhibition hall, restaurant, street, airport and train station; two different channel characteristics are simulated: G712 and MIRS. Aurora 2 offers two training sets – clean set and multi-condition set, and three test sets – test set A, B and C. Matched condition training and testing can be assumed between the multi-condition training set and test set A. As noted in [4], the noises in test set A also seem to cover the spectral characteristics of the noises in test set B. Mismatches exist between the multi-condition training data and test set C, because of the different channel characteristics (G712 versus MIRS). In the experiments, the new UC model was trained based on the clean training set containing 8440 utterances. This model is compared with the two ETSI baseline systems described in [4], trained on the clean training set and multi-condition set respectively. Table 1 presents the results.

Table 1: Aurora 2 word error rates for the ETSI baseline systems and for the new universal compensation (UC) model

<table>
<thead>
<tr>
<th>Model</th>
<th>Training set</th>
<th>Test set</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>ETSI</td>
<td>Clean</td>
<td>38.66</td>
<td>44.26</td>
</tr>
<tr>
<td></td>
<td>Multi</td>
<td>12.19</td>
<td>13.73</td>
</tr>
<tr>
<td>UC</td>
<td>Clean</td>
<td>12.32</td>
<td>13.38</td>
</tr>
</tbody>
</table>

4.2. Aurora 3

Unlike Aurora 2, Aurora 3 consists of digit sequences (in four languages – Finnish, Spanish, German and Danish) recorded in real-world in-car environments, with realistic noise and channel effects. Speech data were recorded in three different noise (driving) conditions – quite, low noise and high noise, and each utterance was recorded simultaneously by using two microphones, a close-talk (CT) microphone and a hand-free (HF) microphone.
Three experimental conditions are defined in Aurora 3: 1) well-matched (WM) condition in which the training and testing sets contain well-matched data for both the microphones and noise conditions; 2) medium-mismatch (MM) condition in which the training and testing data are both from the HF microphones but differ in noise conditions – quite and low-noise data for training and high-noise data for testing; 3) high-mismatch (HM) condition in which the training and testing sets differ in microphones – the data from the CT microphone for all noise conditions for training and the data from the HF microphone for low-noise and high-noise conditions for testing. Since the HF microphone normally picked up more noise from the background, the training and testing sets in the HM condition differ effectively in both microphones and noise levels. In our experiments, the UC model was trained using the training data for the high-mismatch condition, by treating the CT data as “clean” data. This gives 1540, 1696, 1007 and 1720 training utterances, respectively, for Finnish, Spanish, German and Danish.

Table 2 shows the results. The UC model has performed equally well as the baseline model trained for the well-matched conditions. The UC model has outperformed the baseline when mismatches exist between the training and testing conditions. Average relative improvements of 67.44%, 12.68% and 12.35% are obtained for the UC model for the high-mismatch, medium-mismatch and well-matched conditions, respectively.

### Table 2: Aurora 3 word error rates

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>WM (40%)</td>
<td>7.26</td>
<td>7.06</td>
<td>8.80</td>
<td>12.72</td>
<td>8.96</td>
</tr>
<tr>
<td>MM (35%)</td>
<td>19.49</td>
<td>16.69</td>
<td>18.96</td>
<td>32.68</td>
<td>21.96</td>
</tr>
<tr>
<td>HM (25%)</td>
<td>59.47</td>
<td>48.45</td>
<td>26.83</td>
<td>60.63</td>
<td>48.85</td>
</tr>
<tr>
<td>Overall</td>
<td>24.59</td>
<td>20.78</td>
<td>16.86</td>
<td>31.68</td>
<td>23.48</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>WM (40%)</td>
<td>6.05</td>
<td>6.65</td>
<td>7.49</td>
<td>11.19</td>
<td>7.84</td>
</tr>
<tr>
<td>MM (35%)</td>
<td>17.24</td>
<td>15.50</td>
<td>16.27</td>
<td>26.84</td>
<td>18.96</td>
</tr>
<tr>
<td>HM (25%)</td>
<td>12.54</td>
<td>12.17</td>
<td>13.74</td>
<td>19.90</td>
<td>14.58</td>
</tr>
<tr>
<td>Overall</td>
<td>11.58</td>
<td>11.12</td>
<td>12.12</td>
<td>18.84</td>
<td>13.41</td>
</tr>
</tbody>
</table>

### 4.3. More noise conditions unseen in Aurora

Three more noise conditions unseen in Aurora were used to further investigate the ability of the UC model to offer robustness for a wide variety of noises. These three noises are: 1) a mobile phone ringtone, 2) a pop song segment with mixed music and voice of a female singer, and 3) a broadcast news segment for a female singer, and 3) a broadcast news segment.

### 5. Conclusions

A method capable of dealing with arbitrary additive noise based only on clean speech training data and simulated noise data was described. The method, termed universal compensation (UC), has been evaluated on Aurora 2 and 3 and further, on noise conditions unseen in Aurora. The results show that the UC model assuming no knowledge of noise has performed equally well as the baseline model trained for the specific tasks, and has outperformed the baseline when mismatches existed between the training and testing conditions. The UC technique may stand alone, or co-exist with other noise robust techniques to provide improved robustness for applications in which knowledge about the noise or environment is not available.

### 6. References