SPEAKER IDENTIFICATION IN UNKNOWN NOISY CONDITIONS - A UNIVERSAL COMPENSATION APPROACH

Ji Ming[†]*, Darryl Stewart*[†]*, and Saeed Vaseghi*[‡]

[†]School of Computer Science, Queen's University Belfast, Belfast BT7 1NN, UK
[‡]Department of Electronic and Computer Engineering, Brunel University, Middlesex UB8 3PH, UK

ABSTRACT

We consider speaker identification involving background noise, assuming no knowledge about the noise characteristics. A new method, namely *universal compensation* (UC), is studied as a solution to the problem. The UC method is an extension of the missing-feature method, i.e. recognition based only on reliable data but robust to any corruption type, including full corruption that affects all time-frequency components of the speech representation. The UC technique achieves robustness to unknown, full noise corruption through a novel combination of the multi-condition training method and the missing-feature method. Multi-condition training is employed to convert full-band spectral corruption into partial-band spectral corruption, and the missing-feature principle is employed to reduce the effect of the remaining partial-band corruption on recognition by basing the recognition only on the matched or least-distorted spectral components. The combination of these two strategies makes the new method potentially capable of dealing with arbitrary additive noise – with arbitrary temporal-spectral characteristics – based only on clean speech training data and simulated noise data, without requiring knowledge about the actual noise. The SPIDRE database is used for the evaluation, assuming various corruptions from real-world noise data. The results obtained are encouraging.

1. INTRODUCTION

Accurate speaker recognition is made difficult due to a number of factors, with handset/channel mismatch and environmental noise being two of the most prominent. During the past years, much research has been conducted towards reducing the effect of handset/channel mismatch. Linear and nonlinear compensation techniques have been proposed, with applications to feature, model and match-score domains (see, for example, [1]–[4]). In this paper, we study the problem of speaker recognition in noisy conditions, assuming speech samples corrupted by background noise.

To date, research has targeted the impact of background noise through filtering techniques such as spectral subtraction or Kalman filtering [5][6]. Other techniques rely on a statistical model of the noise, for example, PMC (parallel model combination) [7][8]. Recent studies on the missing-feature method suggest that, when knowledge of the noise is insufficient for cleaning up the speech data, one may alternatively ignore the severely corrupted speech data and base the recognition only on the data with least contamination. This can effectively reduce the influence of noise while requiring less knowledge than usually needed for noise filtering or compensation (e.g., [9]–[11]). However, the missing-feature method is only effective for partial noise corruption, i.e., the noise only affects part of the speech representation.

This paper investigates speaker identification involving additive background noise, assuming any corruption type (e.g., full, partial, stationary or time-varying), and furthermore assuming no knowledge about the noise characteristics. The missing-feature method for accommodating partial noise corruption is extended to accommodate *full* noise corruption, i.e. focusing recognition only on least-distorted features while assuming noise, with unknown characteristics, affecting all time-frequency components of the speech representation. The new technique involves a novel combination of the multi-condition training method and 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 only on the matched or least-distorted spectral components. The combination of these two strategies makes the new method potentially capable of dealing with arbitrary additive noise – with arbitrary temporal-spectral characteristics – based only on clean speech training data and simulated noise data, without requiring knowledge about the noise. We term the new technique Universal Compensation (UC). An early study of the model, for robust *speech* recognition, was described in [12].

2. UNIVERSAL COMPENSATION

The UC technique for speaker recognition 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 signal-to-noise ratios (SNRs);
2. Given a test spectrum, search for the components in each model spectrum that best match the corresponding components in the test spectrum, and produce a score based on the matched spectral 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, illustrated in Fig.1, which displays the frequency spectrum of a single frame of speech. Fig.1 shows, on the left, an instance of a clean speech spectrum, representing the data available for training.

This work was supported by the UK EPSRC grant GR/S63236.

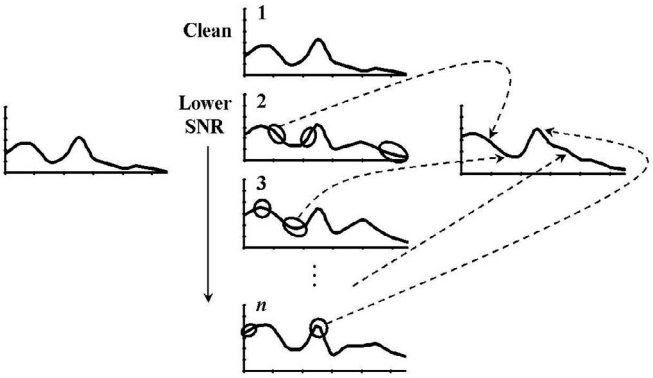

Fig. 1. Illustration of the UC method. Left to right: clean training spectrum, model spectra and noisy test spectrum.

Wide-band flat-spectrum noises (i.e. white noises) with different SNRs are added, respectively, to the waveform of the clean frame, to form a 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 corruption 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, an example of a test spectrum, which is assumed to be the result of the clean frame with the addition of some noise that causes a full-band corruption. The shape 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, model spectra 2, 3 and n, 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, to indicate the probability of the test spectrum associated with the model, i.e. Step 3. Note that a partial-band corruption remains partial in this compensation.

Use of artificially added noise at various SNRs to account for unknown noise sources is not new in speech recognition. The UC method is novel 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, i.e. full, partial, white, colored or none, with simulated noises of a limited variety, e.g. white noise at a limited number of SNRs.

3. ACOUSTIC MODELING FOR RECOGNITION

Formulating the UC method is straightforward following the above example. Assume that L levels of SNR are used 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. Let $p(x|\lambda, l)$ represent a model spectrum, associated with speaker λ and trained for SNR level l ($l = 1, 2, ..., L$), expressed as the probability distribution of the model spectral vector $x = (x_1, x_2, ..., x_N)$ consisting of N components. For convenience, we address the model spectrum by its index (λ, l).

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 estimating the probability of o for each speaker λ, based on the probabilities of o over the individual model spectra (λ, l) associated with λ. As described in Step 2, only the matched components between the test spectrum and the model spectrum are used in the estimation; the mismatched spectral components are ignored to accommodate mismatches between the training and testing conditions. Denote by $o(\lambda, l)$ the matched subset, containing all the matched components in o for model spectrum (λ, l). Given $o(\lambda, l)$ for each (λ, l), the overall probability of o, associated with speaker λ, can be defined as (Step 3):

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

where $p(o(\lambda, l)|\lambda, l)$ is the probability of $o(\lambda, l)$ associated with model spectrum (λ, l), and $p(l|\lambda)$ is a mixture weight, corresponding to the prior probability of SNR level l for speaker λ. In this paper, we assume that the individual spectral components are independent of one another. So the probability $p(o_{sub}|\lambda, l)$ for any subset $o_{sub} \in o$ can be written as

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

where $p(o_n|\lambda, l)$ is the probability for the nth spectral component associated with model spectrum (λ, 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., $o(\lambda, l) = o$). This mixture model including 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 artificial noisy data involving wide-band flat-spectrum noise at 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 mixture density is required in testing for reducing the effect of mismatched noisy spectral components on recognition. To achieve this, we need to decide the matched subset $o(\lambda, l) \in o$ for each model spectrum (λ, l). In principle, the traditional missing-feature methods concerning the the removal of corrupt data based on an estimate of the structure of the corruption may be used to tackle this problem. In this paper,

we consider a solution to the problem by maximizing the appropriate probabilities. If we assume that the matched subset produces a large probability, then $o(\lambda, l)$ may be defined as the subset o_{sub} that maximizes the probability $p(o_{sub}|\lambda, l)$ among all possible subsets in o. However, (2) indicates that $p(o_{sub}|\lambda, l)$ is a function of the size of the subset o_{sub}, implying that the values of $p(o_{sub}|\lambda, l)$ for different sized subsets are of a different order of magnitude and are thus not directly comparable. A solution to this is to replace the conditional probability of the test subset $p(o_{sub}|\lambda, l)$ with the posterior probability of the model spectrum $p(\lambda, l|o_{sub})$, which is defined as follows:

$$p(\lambda, l|o_{sub}) = \frac{p(o_{sub}|\lambda, l)p(\lambda, l)}{\sum_{\lambda', l'} p(o_{sub}|\lambda', l')p(\lambda', l')} \quad (3)$$

where $p(o_{sub}|\lambda, l)$ is the conditional probability as defined in (2) and $p(\lambda, l) = p(l|\lambda)p(\lambda)$ is the prior probability of model spectrum (λ, l), where $p(l|\lambda)$ is defined in (1) and $p(\lambda)$ is the prior probability of speaker λ. The posterior probability $p(\lambda, l|o_{sub})$ 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 $p(\lambda, l|o_{sub})$ with respect to o_{sub}, we should be able to obtain the subset for model spectrum (λ, l) that contains all the matched components in terms of the maximum *a posteriori* (MAP) criterion. Assuming an equal speaker prior $p(\lambda)$, it can be shown that (1) can be expressed in terms of the maximized posterior probability, i.e.

$$p(o|\lambda) \propto \sum_{l=1}^{L} \max_{o_{sub} \in o} p(\lambda, l|o_{sub}) \quad (4)$$

So far we have discussed the calculation of the probability for a single frame. The probability of a speaker given an utterance with T frames $O_1^T = \{o_1, o_2, ..., o_T\}$ can be calculated as

$$p(O_1^T|\lambda) = [\prod_{t=1}^{T} p(o_t|\lambda)]^{1/T} \quad (5)$$

where $p(o_t|\lambda)$ is computed based on (4). Since $p(o_t|\lambda)$ is a probability measure, the normalization in (5) against the length of the observation T makes the value of $p(O_1^T|\lambda)$ also usable as a confidence score to verify the identification result.

4. EXPERIMENTS AND DISCUSSION

The new UC model was evaluated for closed-set speaker identification using the SPIDRE database, a subset of the Switchboard corpus. The database contains 45 target speakers. For each speaker, four conversation halves are provided, of which two are from the same handset. In our experiments, the two conversations from the same handset were used, one for training and the other for testing.

For the UC model, the clean training utterance for each speaker was multiplied by adding simulated wide-band flat-spectrum noise to the utterance at six SNRs: 20 dB, 18 dB, 16 dB, 14 dB, 12 dB and 10 dB. This gives a total of seven training utterances (including the clean training utterance) for each speaker, with seven different SNR levels. Based on these, a UC model, (1), was built for each speaker, with 224 diagonal Gaussian mixtures to account for the expanded training set (on average, 32 mixtures per SNR level). The UC model was compared to a baseline Gaussian mixture model (GMM) with 32 mixtures trained on the clean data. The speech was divided into frames of 20 ms at a frame rate of 10 ms. Each frame was featured using 12 log filter-bank amplitudes, decorrelated by a high-pass filter $H(z) = 1 - z^{-1}$. The first-order delta parameters were appended, thus forming a 24-element feature vector for each frame.

Table 1. Identification accuracy (%) for the new UC model and a baseline GMM, for three different test durations - 15s, 10s and 5s, averaged over six different noises - car, jet engine, mobile phone ring, restaurant, pop song and street

SNR	15s		10s		5s	
(db)	UC	GMM	UC	GMM	UC	GMM
Clean	91.11	91.11	88.89	88.89	86.67	86.67
20	87.78	82.22	88.52	79.63	84.45	76.30
15	85.19	69.26	85.18	68.52	80.00	58.89
10	74.45	41.85	74.82	44.07	68.89	39.26

Table 2. Identification accuracy (%) for different types of noise, averaged over SNR between 10 - 20db

Noise	15s		10s		5s	
	UC	GMM	UC	GMM	UC	GMM
Car	85.93	75.56	82.22	74.07	80.00	68.15
Eng.	86.67	64.44	89.63	64.45	82.22	60.74
Ring	80.00	54.82	78.52	50.37	75.56	40.74
Rest.	76.30	60.74	78.52	61.48	74.07	57.78
Song	87.41	71.85	88.15	74.07	82.22	65.92
Street	78.52	59.26	80.00	60.00	72.59	55.56

Six different real-world noises were used in the experiments. They were: 1) a car noise, 2) a jet engine noise, 3) a polyphonic mobile phone ring, 4) a restaurant babble noise, 5) a pop song segment with mixed music and voice of a female singer, and 6) a street noise. These noises were added, respectively, to each clean test utterance at three SNRs: 20 dB, 15 dB and 10 dB, to simulate real-world noise corruption. The first 5, 10 and 15 seconds of speech from each test conservation were used for test utterances, respectively, corresponding to three different test durations.

Table 1 shows the performances by the UC model and the baseline GMM, as a function of the SNR and test duration, averaged over all the six noises (including the clean condition). Table 2 shows the results from a different angle, giving the results of the two systems for individual noise, averaged over SNR between 10-20db. The two tables indicate that the new UC model has improved over the baseline GMM in all tested noisy conditions, without having assumed any knowledge about the noise. Table 1 also indicates that the UC model is capable of achieving the same identification accuracy as the baseline GMM in clean testing conditions.

For comparison, a multi-condition GMM was trained using both clean and noisy data for car, restaurant and street noise each at three SNRs: 20 dB, 15 dB and 10 dB (thus with a total of ten clean/noisy conditions). The model used 224 mixtures for each speaker (the same number of mixtures as used in the UC model). Fig.2 shows the comparison. The multi-condition model improved over UC in two matched conditions (i.e., car and street). However, the model performed poorer than UC in the restaurant noise and

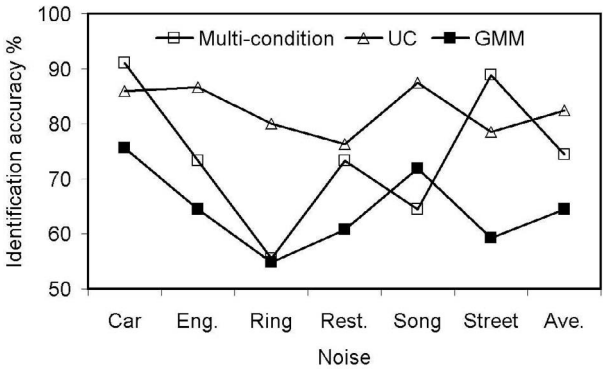

Fig. 2. Comparison to a multi-condition model trained for car, restaurant and street noises and tested in matched/mismatched conditions, averaged over SNR between 10-20db, test duration = 15s.

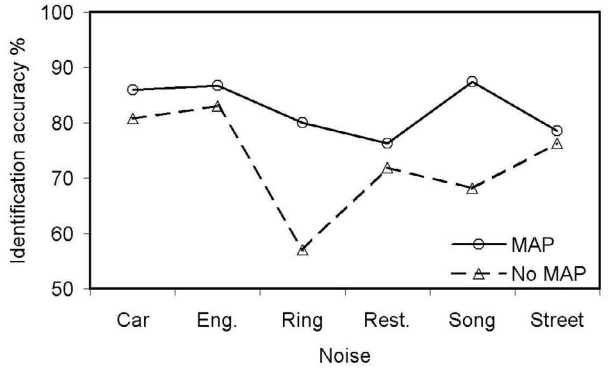

Fig. 3. Difference in performance for different noises, for the UC model with and without the MAP optimization for feature selection, averaged over SNR between 10-20db, test duration = 15s.

in all the unseen noises (i.e., jet engine, phone ring and pop song), resulting in a poorer average performance, indicating the potential of the UC model for dealing with a wider range of noises.

The impact of the MAP optimization for selecting the feature subset in computing the match score, as shown in (4), is illustrated in Fig.3, showing a comparison of performance between the UC models with the optimization and without the optimization. The latter is equivalent to a GMM trained on multi-condition data for wide-band flat-spectrum noise at seven different SNRs, as described above. The optimization has led to improved accuracy in all tested noisy conditions. As expected, the improvement is more significant for the noises that are significantly different in the overall spectral structure from the wide-band flat-spectrum noise used in the compensation. In our experiments, for example, these noises include the mobile phone ring and pop song.

As indicated in (5), the match score produced by the UC model can be used both for identification and additionally as a confidence measure for verification, thereby improving the identification accuracy through rejecting the results with a low confidence. This is illustrated in Fig. 4. For example, at a rejection rate of 30%, it is possible to improve the identification accuracy from about 80% to about 90% for the car noise, and from about 66% to about 77% for the street noise, for an SNR = 10 db and a test duration of 15s.

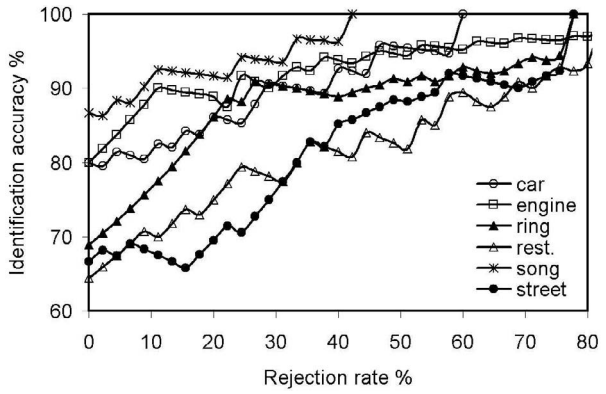

Fig. 4. Identification accuracy as a function of rejection rate, produced by the UC model for different noises. SNR = 10db, test duration = 15s.

5. REFERENCES

[1] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, vol. 10, pp. 19-41, 2000.

[2] L. P. Heck, Y. Konig, M. K. Sonmez M. and Weintraub, "Robustness to telephone handset distortion in speaker recognition by discriminative feature design," Speech Communication, vol. 31, pp. 181-192, 2000.

[3] R. Auckenthaler, M. Carey and H. Lloyd-Thomas, "Score normalisation for text-independent speaker verification systems," Digital Signal Processing, vol. 10, pp. 42-54, 2000.

[4] R. Teunen, B. Shahshahani and L. Heck, "A model-based transformational approach to robust speaker recognition," ICSLP'2000.

[5] J. Ortega-Garcia and L. Gonzalez-Rodriguez, "Overview of speaker enhancement techniques for automatic speaker recognition," ICSLP'96, pp. 929-932.

[6] Suhadi, S. Stan, T. Fingscheidt and C. Beaugeant, " An evaluation of VTS and IMM for speaker verification in noise," Eurospeech'2003, pp. 1669-1672.

[7] T. Matsui, T. Kanno and S. Furui, "Speaker recognition using HMM composition in noisy environments," Computer Speech and Language, vol. 10, pp. 107-116, 1996.

[8] L. P. Wong and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," ICASSP'2001.

[9] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environment with combined spectral subtraction and missing data theory", ICASSP'98, pp. 121-124.

[10] L. Besacier, J. F. Bonastre and C. Fredouille, "Localization and selection of speaker-specific information with statistical modelling", Speech Commun., vol. 31, pp. 89-106, 2000.

[11] J. Ming, D. Stewart, P. Hanna, P. Corr, F. J. Smith and S. Vaseghi, "Robust speaker identification using posterior union models," Eurospeech'2003, pp. 2645-2648.

[12] J. Ming and B. Hou, "Evaluation of universal compensation on Aurora 2 and 3 and beyond," ICSLP'2004.