UNIVERSAL COMPENSATION - AN APPROACH TO NOISY SPEECH RECOGNITION ASSUMING NO KNOWLEDGE OF NOISE

Ji Ming

School of Computer Science
Queen's University Belfast, Belfast BT7 1NN, UK

ABSTRACT

We aim to develop an acoustic model for noisy speech recognition that is "trained once, suits all", in terms of offering a recognition performance close to the matched training-testing condition performance based only on clean speech training data. This paper describes such a method termed *universal compensation*, for its ability to accommodate arbitrary additive noise without assuming any knowledge about the noise. The new UC method consists of two parts: 1) converting full-band spectral corruption into partial-band spectral corruption through compensations for simulated wide-band flat-spectrum noise at consecutive SNRs (signal-to-noise ratios), and 2) reducing the effect of the remaining partial frequency-band corruption on recognition by ignoring the severely mismatched spectral components and basing the recognition mainly on the matched or appropriately compensated spectral components. Experiments on Aurora 2 indicate that the new model, trained from clean data, has achieved a performance comparable to the performance obtained by the baseline system trained on multi-condition data; experiments with noises unseen in Aurora 2 have shown significant improvement for the new model over the baseline model with multi-condition training.

1. INTRODUCTION

Speech recognition performance is known to degrade dramatically when a mismatch occurs between training and testing conditions. The mismatch between training and testing data can be caused by a number of factors, with background noise being one of the most prominent. Traditional approaches for removing the mismatch thereby reducing the effect of noise on recognition include: 1) removing the noise from the testing signal - known as noise filtering or speech enhancement, and 2) constructing a new acoustic model to match the appropriate testing environment - known as noise or environment compensation. Examples of the noise filtering methods include spectral subtraction, Wiener filtering or RASTA filtering (e.g. [1][2]), each assuming the availability of certain knowledge such as the spectral characteristic of the noise. Examples of the noise compensation methods include model adaptation and multi-condition or multi-style training. Model adaptation adjusts a current acoustic model to match a new environment, assuming the availability of training data from the new environment (e.g. [3][4]); multi-condition and multi-style training constructs acoustic models that are suitable for a number of noisy environments, assuming the availability of training data from each of the environments (e.g. [5][6]). More recent studies are focused on the methods requiring less knowledge for the noise or environment, since this can be difficult to obtain in real-world applications involving mobile environments, with unpredictable nonstationary noise. For example, the recently studied missing-feature method (e.g. [7]-[9]) suggests that, when there is a lack of knowledge, we may alternatively detect the severely distorted speech data and subsequently ignore them in the recognition. This can reduce the effect of noise on recognition while requiring less knowledge than usually needed for noise removal or compensation. 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 noise compensation for additive background noise, assuming any corruption type (e.g. full, partial, stationary or time varying), and assuming no knowledge about the noise and no training data from the noisy environment. We propose a new noise compensation method that is capable of accommodating all possible additive distortions, in terms of offering a recognition performance close to the matched training-testing condition performance, based only on clean speech training data. We call the new method Universal Compensation (UC), for its ability to deal with arbitrary additive noise - with arbitrary temporal-spectral characteristic – requiring no knowledge about the noise nor training data from the noisy environment.

2. METHODOLOGY

The UC technique includes three steps [1]:

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 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

[1] Patent is applied for.

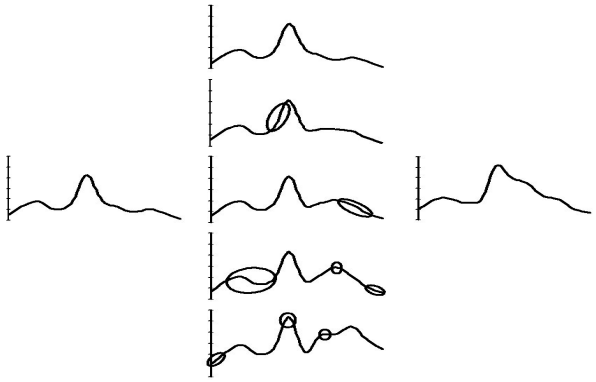

Fig. 1. Illustration of the UC method. Left: a clean spectrum. Middle: model spectra formed by adding different levels of wide-band flat-spectrum noise to the clean waveform (SNR decreases from top to bottom). Right: a noisy test spectrum. The matched components between the test spectrum and each of the model spectra are enclosed within the circles on the appropriate model spectra.

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 mode spectra, for example, 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. The effect of partial-band corruption on recognition can be reduced by ignoring the distorted spectral components and by basing the score only on the matched or least distorted spectral components, i.e. Step 2. 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.

The accuracy of the method for converting a full-band corruption into partial-band corruption is determined by two factors: the frequency-band resolution and the amplitude resolution. The band resolution determines the bandwidth for each spectral component. The smaller this bandwidth, the more accurate the approximation for an arbitrary noise spectrum by a flat spectrum located in the same frequency band. The amplitude resolution refers to the quantizing steps for the SNR, used to generate the wide-band flat-spectrum distortion. Given the range of SNR, the finer the quantizing steps, the more accurate the approximation for any level of noise. Given a test spectrum, there may be some model spectra without matched components. These model spectra can be ignored in Step 3 assuming correspondingly low probabilities. Note that a partial-band corruption remains partial in the conversion.

3. FORMULATION

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 $p(x \mid s, l)$ the probability distribution for a model spectrum associated with speech state s and trained for SNR level l, $l = 1, 2, ..., L$.

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 state s, based on the matching components between the test spectrum and each of the model spectra associated with state s. Denote by $o(s, l)$ the subset in o containing all the matching components for model spectrum $p(x \mid s, l)$, addressed by (s, l). Both the size and components of $o(s, l)$ can be different 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

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

where $p(o(s, l) \mid s, l)$ is the probability of o associated with model spectrum (s, l), 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 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 $p(o_{sub} \mid s, l)$ for any subset $o_{sub} \in o$ can be written as

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

where $p(x_n \mid s, l)$ is the probability distribution of the nth spectral component 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. 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, could 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 $o(s, l)$ may be defined as the subset o_{sub} that maximizes the probability $p(o_{sub} \mid s, l)$ among all possible subsets in o. However, (2) indicates that the value of $p(o_{sub} \mid s, l)$ is a function of the size of

the subset o_{sub}, implying that the values of $p(o_{sub} \mid s, l)$ for different sized subsets are of a different order of magnitude and are thus not directly comparable. A possible solution to this is to replace the conditional probability of the test subset, $p(o_{sub} \mid s, l)$, with the posterior probability of the model spectrum, $p(s, l \mid o_{sub})$. The posterior probability of model spectrum (s, l) given a test subset o_{sub} is defined as

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

where $p(o_{sub} \mid s, l)$ 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 $p(s, l \mid o_{sub})$ 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 $p(s, l \mid o_{sub})$ with respect to o_{sub}, we should be able to obtain the subset $o(s, l)$ 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:

$$o(s, l) = \arg \max_{o_{sub} \in o} p(s, l \mid o_{sub}) \quad (4)$$

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

$$p(o \mid s) \propto \sum_{l=1}^{L} w(s, l) \max_{o_{sub} \in o} p(s, l \mid o_{sub}) \quad (5)$$

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

4. EXPERIMENTAL EVALUATION

Aurora 2 is used to evaluate the performance of the new method. The new method is incorporated into an HMM and trained using only the clean training set. 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=20dB, reducing 2dB every level, until SNR=2dB. The wide-band flat-spectrum noise is computer-generated white noise filtered by a low-pass filter with a 3dB-bandwidth of 3.5kHz.

The speech is divided into frames of 25ms at a frame rate of 10ms. For each frame, we use a 12-channel mel-frequency filter bank to estimate 12 log spectral energies (i.e. log FB energies). These 12 log FB energies are decorrelated by using a decorrelation filter $H(z) = 1 - z^{-1}$, and are then grouped uniformly into six subbands. For each decorrelated log FB energy, its delta and delta-delta coefficients are also calculated and grouped into six subbands in the same way as for the static spectral coefficients. Thus, for each frame, we have a spectral vector consisting of a total of 18 components, six for the static spectral components, six for the delta spectral components and six for the delta-delta spectral components, for the six subbands. Each component contains two coefficients, and so the overall size of the spectral vector for a frame is 36. In the experiments, each digit is modeled by 15 states and each state is modeled with 32 mixtures, accounting for the expanded training set including both the clean data and the artificial noisy data with ten SNR levels. Each mixture component is a Gaussian density with a diagonal covariance matrix. The performance of the new UC model is compared with the performance of the baseline system defined by ETSI, presented in [5].

4.1. Tests on Aurora Conditions

A performance measure for a whole test set, as average word accuracy over all noises and over SNRs between 0 and 20dB, is introduced in [5]. This measure is used to compare the results. Two baseline systems are described in [5], one trained on the clean training set and the other on the multi-condition training set. As described, the new UC model is trained effectively using only the clean training set from the database.

Table 1 shows the average performance on test set A, obtained by the new UC model, compared with the results by the two ETSI baseline systems. Fig. 2 further shows the results as a function of SNR, averaged over all noises (including clean condition). Table 1 and Fig. 2 for test set A indicates that the new model has significantly improved over the baseline system trained on clean data and tested in noisy conditions, and that the new model has achieved an accuracy comparable to the matched training-testing condition accuracy without having assumed any knowledge about the noise.

Next, Table 2 shows the recognition results on test set B, for the same systems used for test set A. The results as a function of SNR averaged over all noises are also included in Fig. 2. The similarity of the noise characteristics between test set A and test set B is indicated by the similarity of the average performance between Table 2 and 1, and between the corresponding curves in Fig. 2. As shown in Table 2 and Fig. 2, the new UC model has offered slightly better average performance than the baseline system trained on the multi-condition data.

Finally, Table 3 shows the results on test set C, for the same systems used above. The results as a function of SNR averaged over the noises are also shown in Fig. 2. Comparing Table 3 with Table 1 and 2, it is seen that the baseline system trained on multi-condition data has experienced performance degradation on test set C (e.g., the average accuracy dropped from 87.81% for test set A to 83.77%), due to the mismatched channel characteristics (i.e. MIRS versus G712). The new UC model, searching for the matched components not only between the static spectra but also between the channel-insensitive dynamic spectra, has coped with this mismatch more effectively, offering an accuracy of 87.30% for test set C that is close to the matched condition performance (87.81%) for test set A. Fig. 2 indicates that the improvement for the new model is more significant for low SNR conditions.

Table 1. Word accuracy (%) on test set A, averaged over SNRs between 0 - 20dB, for the new universal compensation (UC) model, compared with the ETSI baseline systems

Model	Training set	Noise condition				Ave.
		Sub.	Bab.	Car	Exhib.	
UC	Clean	88.01	85.36	90.48	86.87	87.68
ETSI	Clean	69.48	49.88	60.60	65.39	61.34
ETSI	Multi	88.75	87.95	86.52	88.03	87.81

Table 2. Word accuracy (%) on test set B

Model	Training set	Noise condition				Ave.
		Rest.	St.	Air.	Sta.	
UC	Clean	84.49	88.00	86.89	87.09	86.62
ETSI	Clean	52.59	61.51	53.25	55.63	55.74
ETSI	Multi	85.39	87.03	87.64	85.01	86.27

Table 3. Word accuracy (%) on test set C

Model	Training set	Noise condition		Average
		Subway	Street	
UC	Clean	87.45	87.16	87.30
ETSI	Clean	66.16	66.11	66.14
ETSI	Multi	83.24	84.31	83.77

4.2. More Noise Types

Two more noise conditions unseen in Aurora 2 are used to evaluate the new UC model and to compare its performance with the performance of the baseline system trained on the Aurora multi-condition data. The purpose of these tests is to further investigate the ability of the new model to offer robust performance for a wide variety of noises, i.e., the ability of "trained once, suits all". These two noises are: 1) a mobile phone ringtone, and 2) a pop song segment with a mixture of background music and the voice of a female singer. The spectral characteristics of the two noises are shown in Fig. 3. Table 4 presents the recognition results. As indicated in the table, the new model has offered significantly improved accuracy over the baseline model.

5. SUMMARY

A method capable of dealing with arbitrary additive noise based only on clean speech training data is described. The new method, termed universal compensation (UC), has been evaluated on Aurora 2. Trained using information from the clean training set, the new method has achieved a performance comparable to the performance obtained by the baseline system trained on multi-condition data. Further experiments with noises unseen in Aurora 2 have shown significant improvement for the new model over the baseline model with multi-condition training.

Table 4. Word accuracy (%) with two noises, a mobile phone ringtone and a pop song, unseen in Aurora 2, for the new UC model, compared with the ETSI baseline system

SNR (dB)	Model	Training set	Noise condition		Ave.
			Ringtone	Pop song	
10	UC	Clean	95.43	87.47	91.45
	ETSI	Multi	76.60	76.57	76.58
5	UC	Clean	92.82	79.18	86.00
	ETSI	Multi	64.29	63.16	63.72
0	UC	Clean	90.11	65.27	77.69
	ETSI	Multi	52.50	44.77	48.63

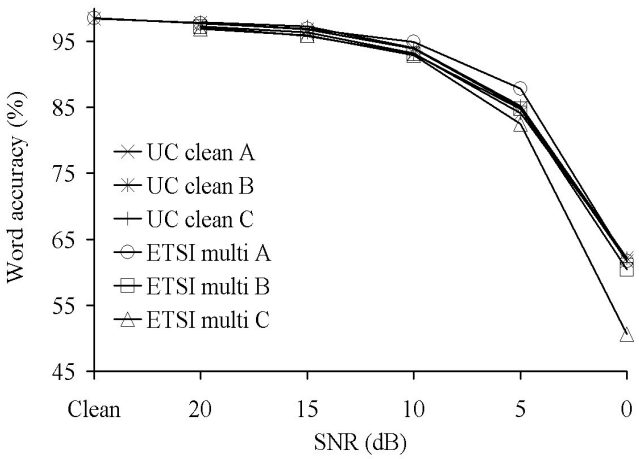

Fig. 2. Word accuracy as a function of SNR for test set A, B and C, for the new UC model trained from clean data, and for the ETSI baseline model trained on multi-condition data.

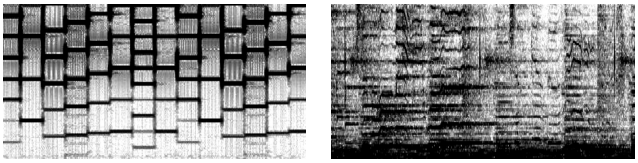

Fig. 3. Spectra of a mobile phone ringtone (left) and a pop song (right) unseen in Aurora 2

6. REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech. *Proceedings of IEEE*, vol. 67, 1586-1604, 1979.

[2] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 578-589, 1994.

[3] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Comput. Speech Language*, vol. 10, pp. 249-264, 1996.

[4] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech and Audio processing*, vol. 4, pp. 352-359, 1996.

[5] D. Pearce and H.-G. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *ISCA ITRW ASR2000*.

[6] L. Deng, A. Acero, M. Plumpe and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," *ICSLP'2000*, pp. 806-809.

[7] R. P. Lippmann and B. A. Carlson, "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise," *Eurospeech'97*, pp. 37-40.

[8] M. Cooke, P. Green, L. Josifovski and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data." *Speech Communication*, vol. 34, pp. 267-285, 2001.

[9] J. Ming, P. Jancovic and F. J. Smith, "Robust speech recognition using probabilistic union models," *IEEE Trans. Speech Audio Processing*, vol. 10, pp.403-414, 2002.