INCORPORATING MASK MODELLING FOR NOISE-ROBUST AUTOMATIC SPEECH RECOGNITION

Münevver Köküer and Peter Jančovič

School of Electronic, Electrical & Computer Engineering, University of Birmingham, Birmingham, UK
{m.kokuer, p.jancovic}@bham.ac.uk

ABSTRACT

In this paper we investigate an incorporation of mask modelling into an HMM-based ASR system. The mask model is estimated for each HMM state and mixture by using a separate Viterbi-style training procedure and it expresses which regions of the spectrum are expected to be uncorrupted by noise for the HMM state. Experimental evaluation is performed on noisy speech data from the Aurora 2 database. Significant performance improvements are achieved when the mask modelling is incorporated within the standard model and two models that had already compensated for the effect of the noise.

Index Terms— automatic speech recognition, mask modelling, noise robustness, missing-feature theory.

1. INTRODUCTION

The performance of automatic speech recognition (ASR) systems degrades rapidly when speech signal is corrupted by a background acoustical noise. There have been several different ways to improving noise robustness. Feature representation of the speech signal that is more robust to the effect of noise can be sought. Speech signal can be enhanced prior to its employment in the recogniser by techniques such as spectral subtraction, Wiener filtering, or MAP-based enhancement, e.g., [1] [2]. Assuming availability of some knowledge about the noise, noise-compensation techniques, e.g., [3], can be applied in the feature or model domain to reduce the mismatch between the training and testing data. Considering that only information on the location of noise-corrupted spectro-temporal elements is available (referred to as mask), the missing feature theory (MFT) can be employed for improving noise robustness [4] [5] by marginalising these elements in the observation probability calculation.

Recently, several techniques has been proposed which aim to exploit the speech signal properties, such as, the spectral peaks being more robust to a broad-band noise than the spectral valleys or harmonicity information. The authors in [6] proposed a technique that performs locking of the spectral peak-to-valley ratio in order to alleviate the mismatch between clean and noisy features caused by the spectral valleys being buried by noise. The authors in [7] [8] appended the information on spectral peaks into the acoustic features. This was shown to improve the recognition accuracy on clean speech and isolated-word recognition in noisy conditions. The authors in [9] modified the likelihood calculation with the aim of emphasising parts of the spectrum corresponding to peaks.

In this paper, we investigate an incorporation of the mask modelling into an HMM-based automatic speech recognition (ASR) system in noisy conditions. As the mask expresses which spectro-temporal regions are uncorrupted by noise, the proposed technique can also be seen as a generalised and soft incorporation of the spectral peak information. We have introduced this technique in [10], where evaluations were performed on intervocalic English consonant recognition task. In the proposed model, the mask model is associated with each HMM state and mixture and it expresses what mask information the state/mixture would expect to find in the signal. The mask modelling is performed by employing the Bernoulli distribution whose parameters are estimated by a separate Viterbi-style training procedure after the HMMs are trained using the acoustic features. The incorporation of the mask modelling is evaluated in a standard model and in two models that had compensated for the effect of the noise, missing-feature and multicondition training model. Experiments are performed on the Aurora 2 database. Experimental results show significant improvements in recognition performance in strong noisy conditions achieved by the models incorporating the mask modelling.

2. INCORPORATING MASK MODELLING INTO HMM-BASED ASR SYSTEM

Let Y be the sequence of observation vectors extracted from a given speech utterance. The goal of a speech recogniser is to find the word sequence $\hat{W}$ that maximises the posterior probability $P(W|Y)$. Let us consider that an additional information reflecting which spectro-temporal region contains an uncorrupted information is available and this is contained in the sequence of mask vectors M. Considering an HMM-

based ASR system with the proposed incorporation of mask modelling, the search for $\hat{W}$ can then be expressed as

$$\hat{W} \approx \arg\max_{W} P(Y|M,S,W)P(M|S,W)P(S|W)P(W) \quad (1)$$

where $P(S|W)$ is the HMM state-transition probability, S is the sequence of HMM states used during the recognition, and $P(W)$ is the language-model probability. The term $P(Y|M,S,W)$ is the probability of the observation sequence Y, which is (unlike in the standard model) now conditioned also on the mask M – as such, this term corresponds to the employment of the missing-feature technique.

The term $P(M|S,W)$ is referred to as mask-model probability and its incorporation is the novelty presented in this paper. This term expresses how likely the given mask M is being generated by the HMM state sequence S. The mask-model probability $P(M|S,W)$ serves as a penalisation factor for states whose mask model is not in agreement with the mask extracted from the given acoustic signal.

Having an example of noise (or knowledge of noise characteristics), the mask model could be estimated based on masks obtained from the training data corrupted by the given noise. Having no information about noise, it could be estimated by using a mask reflecting some a-priori knowledge about speech, for instance, the fact that high-energy regions of speech spectra are less likely to be corrupted by noise. In this paper, the training of the mask model was performed by employing the oracle masks estimated on the training data corrupted by various noises and at various SNR levels (the multicondition training data).

The estimation of the mask model is performed by a separate training procedure that is performed after the HMMs have been trained (i.e., the trained HMMs are not altered). The following sections give detailed description of the estimation of the mask model and its incorporation during the recognition.

2.1. Estimating the mask model for HMM states

Let $\mathbf{m} = \{m(1),\ldots,m(B)\}$ denotes the mask vector at a given frame, where $m(b)$ is the binary mask information of the channel b and B is the number of channels. We model the mask-model probability $P(\mathbf{m}|l,s)$ for each HMM state s and mixture l using the multivariate Bernoulli distribution as

$$P(\mathbf{m}|l,s) = \prod_{b=1}^{B} \mu_{b,l,s}^{m(b)} (1-\mu_{b,l,s})^{1-m(b)} \quad (2)$$

where $\mu_{b,l,s}$ is the parameter of the distribution. The estimation of the parameters $\mu_{b,l,s}$ of the mask models at each HMM state and mixture can be performed by a Baum-Welch or Viterbi -style training procedure; the latter was used in this paper.

Given a speech utterance, we have a sequence of feature vectors $Y = \{\mathbf{y}_1,\ldots,\mathbf{y}_T\}$ and the corresponding sequence of mask vectors $M = \{\mathbf{m}_1,\ldots,\mathbf{m}_T\}$ where T is the number of frames. The Viterbi algorithm is then used to obtain the state-time alignment of the sequence of feature vectors on the HMMs corresponding to the speech utterance. This provides an association of each feature vector $\mathbf{y}_t$ to some HMM state s. The posterior probability that the mixture-component l (at the state s) have generated the feature vector $\mathbf{y}_t$ is then calculated as

$$P(l|\mathbf{y}_t,s) = \frac{P(\mathbf{y}_t|s,l)P(l|s)}{\sum_{l'} P(\mathbf{y}_t|s,l')P(l'|s)} \quad (3)$$

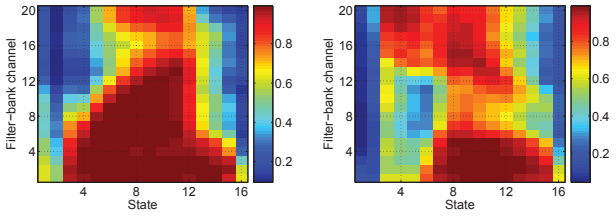
where the mixture-weight $P(l|s)$ and the probability density function of the features used to calculate the $P(\mathbf{y}_t|s,l)$, are obtained as an outcome of the HMM training.

For each mixture l and HMM state s, we collect (over the entire training data-set) the posterior probabilities $P(l|\mathbf{y}_t,s)$ for all $\mathbf{y}_t$'s associated with the state s together with the corresponding mask vectors $\mathbf{m}_t$'s. The parameters $\mu_{b,l,s}$ of the mask models are then estimated as

$$\mu_{b,l,s} = \frac{\sum_{t:\mathbf{y}_t \in s} P(l|\mathbf{y}_t,s)\cdot m_t(b)}{\sum_{t:\mathbf{y}_t \in s} P(l|\mathbf{y}_t,s)} \quad (4)$$

where $m_t(b)$ is the binary mask value.

Examples of the estimated mask model parameters for HMMs of digits 'one' and 'two' are depicted in Figure 1. Regions of a high value of the mask model parameter reflect that the masks associated with the given state were for those regions often one, i.e., little affected by noise. For instance, it can be seen that in digit 'two' the states from three to five (which are likely to correspond to phoneme /t/) have high values of the parameter in high frequency regions.



(a) (b)

Fig. 1. *Examples of the estimated mask model parameters for HMMs of digits 'one' (a) and 'two' (b).*

2.2. Mask-probability incorporation during recognition

The value of the mask-probability when being incorporated in the overall probability calculation in Eq. 1 may need to be scaled in order to achieve an appropriate effect of the mask-model probability on the overall probability (akin to language-model scaling). This can be performed by employing a sigmoid function to transform the $P(m(b)|s,l)$ for each b to a new value, i.e.,

$$P(m(b)|s,l) = \frac{1}{1+e^{-\alpha(P(m(b)|s,l)-0.5)}} \quad (5)$$

where α is a constant defining the slope of the function and the value 0.5 gives shift of the function. The bigger the value of α is the greater the effect of the mask-probability on the overall probability. An appropriate value for α can be decided based on a small set of experiments on a development data.

3. EXPERIMENTAL EVALUATIONS

The experiments were carried out on the Aurora 2 English language connected-digit database. The frequency-filtered logarithm filter-bank energies [11] were used as speech feature representation, due to their suitability for missing-feature based recognition. These were obtained with the following parameter set-up: frames of 32 ms length with a shift of 10 ms between frames were used; both preemphasis and Hamming window were applied to each frame; the short-time magnitude spectra, obtained by applying the FFT, was passed to Mel-spaced filter-bank analysis with 20 channels; the obtained logarithm filter-bank energies were filtered by using the filter $H(z)=z-z^{-1}$ [11]. A feature vector consisting of 18 elements was obtained (the edge values were excluded). In order to include dynamic spectral information, the first-order delta parameters were added to the static FF-feature vector. A continuous-observation left-to-right HMM with 16 states (no skip allowed) was used to model each digit; the pdf at each state was modelled with three and ten Gaussian mixtures when using clean and multicondition training, respectively, and diagonal covariance matrices. The training of HMMs was performed on utterances from the training set. The noisy speech data from the Set A in Aurora 2 were used for recognition experiments. The experimental evaluations were performed by using an in-house speech recognition system for training the mask models and the Hidden Markov Model Toolkit (HTK) [12], which was modified to include the missing-feature method and mask modelling.

The evaluation of the proposed incorporation of mask modelling is first performed by employing the oracle mask. Recognition results obtained for the standard model are presented in Figure 2. Evaluations were also performed on two types of models that had compensated for the effect of noise in order to determine whether incorporating the mask modelling can still provide improvements (as noise compensation could decrease the amount of disagreement between the current mask information and mask models). Results are presented for the missing-feature model (marginalisation employing the oracle mask in order to provide an idealised noise-compensation) in Figure 3 and for the multicondition trained model in Figure 4. It can be seen that incorporating the mask modelling provides significant recognition accuracy improvements in all noisy conditions on the standard model and both models that had already compensated for the noise. These results demonstrate that having an accurately estimated mask (i.e., mask close to the oracle mask), the incorporation of the mask modelling can provide significant improvements.

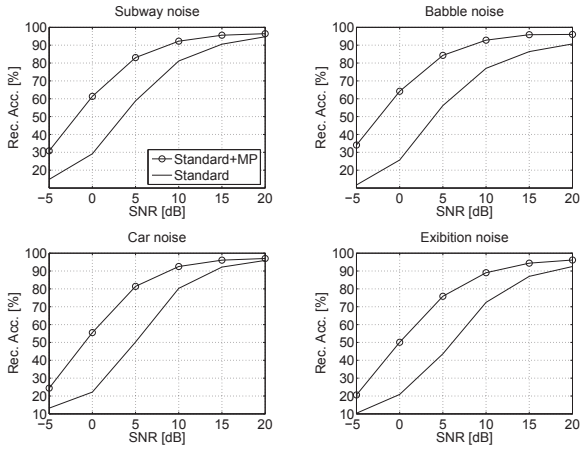


Fig. 2. *Recognition accuracy results obtained by the standard model without and with incorporated mask-probability.*

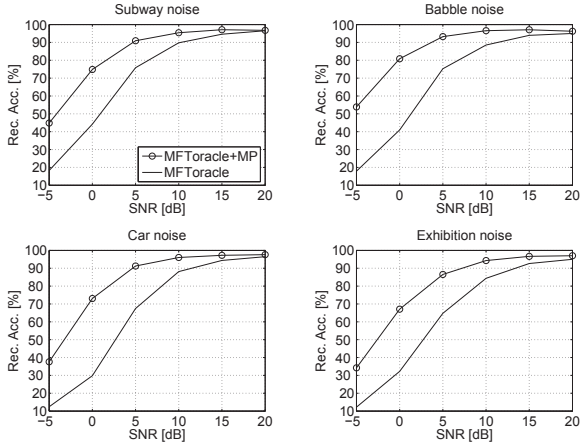


Fig. 3. *Recognition accuracy results obtained by the MFT model without and with incorporated mask-probability.*

We also present experimental results obtained when employing an estimated mask. As the mask estimation is not the focus of this paper a simple mask estimation procedure based on a noise-estimate and sub-band voicing information (obtained by technique presented in [13]) were employed for estimation of the uncorrupted unvoiced and voiced regions, respectively. The results obtained when the estimated mask was employed in the MFT-based system are presented in Figure 5. It can be seen that the mask estimation procedure is not very accurate as only moderate improvements of the MFT are obtained over the standard model. Despite of this, the incorporation of the mask modelling provides still considerable recognition accuracy improvements.

4. CONCLUSION

In this paper, we presented an incorporation of mask modelling into HMM-based ASR system. The mask model was

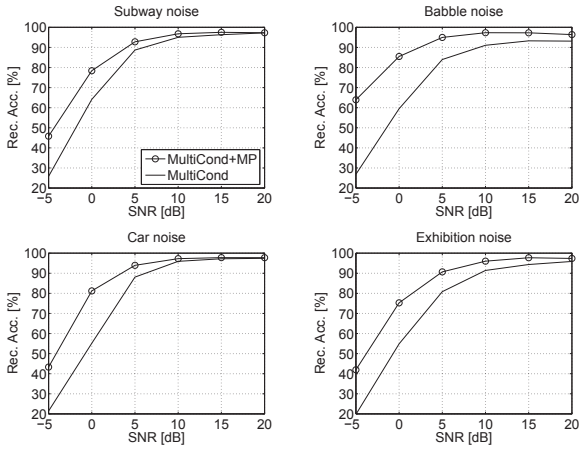


Fig. 4. *Recognition accuracy results obtained by the multi-condition trained model without and with incorporated mask-probability.*

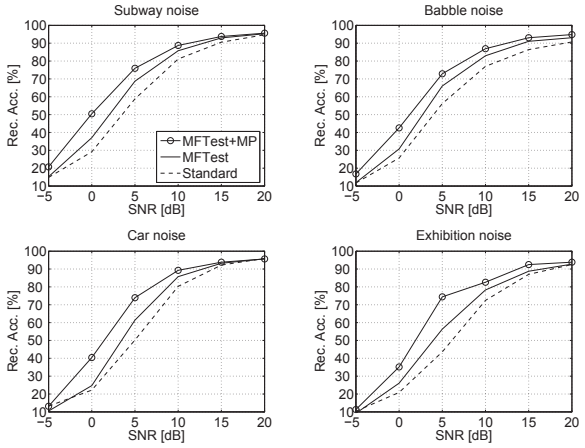


Fig. 5. *Recognition accuracy results obtained by the MFT model using an estimated mask without and with incorporated mask-probability.*

estimated by a separate training procedure for each mixture at each HMM state. The effectiveness of the method was demonstrated within a standard model and two types of noise-compensated models, missing-feature and multicondition training. Experimental evaluations were performed on noisy speech data from the Aurora 2 database. Employing the oracle masks, significant performance improvements in all noisy conditions were obtained when the mask modelling is incorporated in the standard model, and also in both models which had already compensated for the effect of noise. Evaluations were also performed employing an estimated mask obtained by a simple method on the MFT-based noise-compensated model and considerable performance improvements were observed.

This work was supported by UK EPSRC grants EP/D033659/1 and EP/F036132/1.

5. REFERENCES

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustic, Speech, and Signal Proc.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[2] S.V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, John Wiley & Sons, 2005.

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

[4] 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, Rhodes, Greece*, pp. 37–40, 1997.

[5] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, 2001.

[6] Q. Zhu and A. Alwan, "Non-linear feature extraction for robust speech recognition in stationary and non-stationary noise," *Computer Speech and Language*, vol. 17, pp. 381–402, 2003.

[7] M. Padmanabhan, "Spectral peak tracking and its use in speech recognition," *ICSLP, Beijing, China*, 2000.

[8] G. Farahani, S.M. Ahadi, and M.M. Homayounpoor, "Use of spectral peaks in autocorrelation and group delay domains for robust speech recognition," *ICASSP, Toulouse, France*, vol. I, pp. 517–520, 2006.

[9] C. Huang, Y. Huang, F. Soong, and J. Zhou, "Weighted likelihood ratio (WLR) hidden Markov model for noisy speech recognition," *ICASSP, Toulouse, France*, vol. I, pp. 37–40, 2006.

[10] P. Jančovič and M. Köküer, "On the mask modeling and feature representation in the missing-feature ASR: Evaluation on the Consonant Challenge," *Interspeech, Brisbane, Australia*, pp. 1777–1780, 2008.

[11] C. Nadeu, D. Macho, and J. Hernando, "Time and frequency filtering of filter-bank energies for robust HMM speech recognition," *Speech Communication*, vol. 34, pp. 93–114, 2001.

[12] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book. V2.2*, 1999.

[13] P. Jančovič and M. Köküer, "Estimation of voicing-character of speech spectra based on spectral shape," *IEEE Signal Processing Letters*, vol. 14, no. 1, pp. 66–69, Jan. 2007.