Robust speech recognition from binary masks

Arun Narayanan[a)]
Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210
narayaar@cse.ohio-state.edu

DeLiang Wang
Department of Computer Science and Engineering, and Center for Cognitive Science, The Ohio State University, Columbus, Ohio 43210
dwang@cse.ohio-state.edu

Abstract: Inspired by recent evidence that a binary pattern may provide sufficient information for human speech recognition, this letter proposes a fundamentally different approach to robust automatic speech recognition. Specifically, recognition is performed by classifying binary masks corresponding to a word utterance. The proposed method is evaluated using a subset of the TIDigits corpus to perform isolated digit recognition. Despite dramatic reduction of speech information encoded in a binary mask, the proposed system performs surprisingly well. The system is compared with a traditional HMM based approach and is shown to perform well under low SNR conditions.

PACS numbers: 43.72.Ne, 43.72.Bs [DOS]
Date Received: August 16, 2010 **Date Accepted:** September 15, 2010

1. Introduction

Robustness is one of the most important challenges facing automatic speech recognition (ASR) today. Traditional methods perform well under clean speech conditions, but suffer from large performance degradation under noisy environments. The mismatch in training and testing/deployment conditions essentially causes the performance degradation, and is currently handled in many ways. Some of the approaches extract noise robust features, for example RASTA and cepstral mean normalization. In source driven approaches, a speech enhancement algorithm (e.g., Ephraim and Malah, 1985) is applied to the noisy speech and then recognition is performed on the enhanced noisy speech using clean speech models. If noise samples are available *a priori*, noise models may be trained and recognition performed using trained speech and noise models. However, performance of the aforementioned approaches is inadequate in real environments. Robustness of human listeners has been attributed to the human ability of auditory scene analysis (Bregman, 1990). ASR methods coupled with computational auditory scene analysis (CASA) include missing data and uncertainty transformation techniques (Cooke *et al.*, 2001; Srinivasan and Wang, 2007).

All these methods make extensive use of speech features, either in cepstral or spectral domain. A recent study in speech perception shows that the pattern of an Ideal Binary Mask (IBM) appears to provide sufficient information for human speech recognition (Wang *et al.*, 2008). In this study, ideal binary masks are used to modulate speech shaped noise (SSN), which is a stationary noise with a long-term spectrum matching that of natural speech. Human subjects then listen to IBM-gated noise and, despite a dramatic reduction of speech information, are able to recognize speech almost perfectly. The study suggests that IBMs encode phonetic information for humans to perform speech recognition.

[a)] Author to whom correspondence should be addressed.

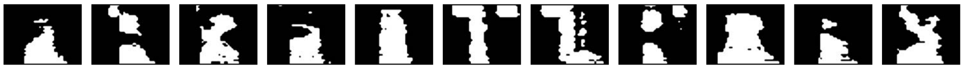

Fig. 1. Typical ideal binary masks of isolated digit utterances 1–9, 'oh' and 'zero', ordered from left to right. In the figure, a white pixel indicates 1 and black pixel 0.

Does a binary pattern provide an adequate basis for ASR? In the current study, we explore this question by performing automatic speech recognition directly on binary patterns of IBM and their variants. We emphasize that this study represents a simple but radically different approach to robust speech recognition.

2. System description

Our ASR system is a binary pattern classifier which classifies ideal binary masks created using isolated digit samples from the TIDigits corpus (Leonard, 1984). IBM is a time-frequency mask, which is a 2D matrix of binary values. An entry in the matrix assumes the value 1 if the corresponding T-F unit has an SNR that exceeds a threshold termed the local SNR criterion (LC). Figure 1 shows typical IBMs for digits 1–9, 'oh' and 'zero' created for 6 dB mixtures of clean speech and SSN, using a LC of 0 dB. As can be seen, the binary patterns of the IBM for these 11 utterances are discernible to the human eye. This encourages the use of ideal binary masks for the task of automatic speech recognition.

The IBMs are created by using the pre-mixed signals. The clean speech and the noise signal, scaled to a specific SNR (signal-to-noise ratio), are first passed though a 64-channel gammatone filter bank with center frequencies spaced according to the ERB (Equivalent Rectangular Bandwidth) scale. Each filter response is then windowed into time frames using a 20-ms rectangular window and a frame shift of 10 ms, to produce a cochleagram (Wang and Brown, 2006). An IBM is then created by calculating the local SNR within each T-F (Time-Frequency) unit and comparing it with the LC. To have the same size for all patterns for recognition, a small window of 64 contiguous frames is selected from the IBM such that the binary pattern is centered in the selected window. This is done by calculating the centroid of the IBM pattern and choosing 32 frames in either direction. The selected window can be thought of as a bounding box enclosing the IBM pattern.

Handwritten digit recognition is a similar binary pattern classification task, which can be performed with very high recognition rates. Convolutional neural networks (CNN) have been widely used for the task of isolated handwritten digit recognition with considerable success (LeCun *et al.*, 1998; Simard *et al.*, 2003). A CNN is suited to capture spatial topology and provide some degree of invariance to translation and size of the input pattern. Since the binary patterns of IBM are, in a way, similar to handwritten digits, we used a CNN to perform the task of digit recognition. The architecture of our CNN, which is shown in Fig. 2, is similar to a LeNet5 CNN described by LeCun *et al.* (1998). Layers C1, C3, and C5 are convolutional layers with the weight kernels of size 5×5, 6×6 and 5×5, respectively. S2 and S4 are subsampling layers with the kernels of size 3×3. The final layer is a fully connected network with 11 output nodes. According to Simard *et al.* (2003), the number of nodes, layers and feature maps does not critically affect the performance as long as there are enough of them. Hence we chose 7, 20 and

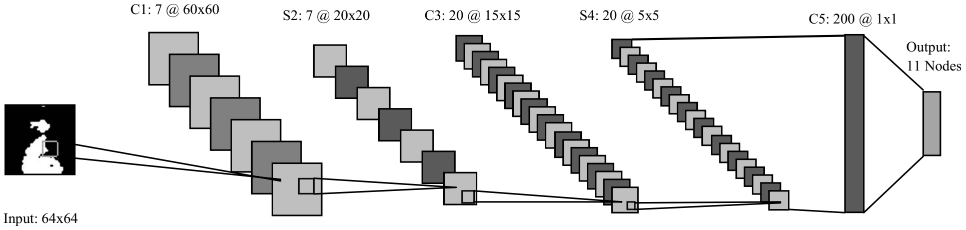

Fig. 2. Architecture of a convolutional neural network for isolated digit recognition.

200 feature maps in the convolutional layers C1, C3 and C5, respectively. LeNet5 uses fewer feature maps, but the size of the input image is smaller than the size of the binary pattern we use (32 × 32 as compared to 64 × 64). The number of nodes in each feature map and the specifications of each subsampling layer are dictated by the size of input pattern and the weight kernel of the preceding layer. All nodes have a hyperbolic tangent activation function. The error function measures the mean squared error. The network is trained using a fast stochastic diagonal Levenberg-Marquardt method for 15 iterations (Lecun *et al.*, 1998).

3. Results

3.1 Experimental setup

As mentioned in the previous section, a subset of the TIDigits corpus was used to evaluate the proposed system. The training and test set consisted of isolated digit utterances from 55 and 56 male speakers, respectively. The speakers in the training set and the test set were different. Each speaker had 2 utterances per digit. The IBM patterns for training were created from 6 dB mixtures, with the LC of 0 dB. These values were chosen for two reasons. First, for a 0 dB mixture, a LC of −6 dB was found to be suitable for human listeners (Wang *et al.*, 2009). Second, if we co-vary the mixture SNR and the LC, the IBM remains the same. Hence, for a 6 dB mixture, a LC of 0 dB will produce IBMs that would have been suitable for human listeners. The training set was created using three noise types: SSN, 32-talker babble noise and party noise. Seven test sets were created for each of the noise types to test the robustness of the system to variations from the training condition. Test sets were created at −6, −3, 0, 3, 6, 9, and 12 dB mixture SNR while fixing the LC to 0 dB in all the cases.

IBMs can be created only if we have access to premixed signals. In real situations IBMs have to be estimated directly from the noisy speech. To gauge how well the proposed method works in real situations, IBMs were also estimated using a CASA based system. The system we used to estimate the IBM is a two stage system as described by Hu and Wang (2009). In the first stage it estimates a voiced mask using MLPs (multilayer perceptrons) trained on pitch based features. The threshold of the MLP was set to −0.2 during testing, as it was found to produce better masks for the training set as far as recognition was concerned. In the second stage, it uses the computed voiced mask to estimate the noise energy and the mask in the unvoiced intervals by calculating the local SNR at each T-F unit and comparing it with the LC. To match the training condition, the LC was set to mixture SNR minus 6 dB.

The CNN was trained using the IBM patterns from the training set. During testing, a window of 64 frames is selected after calculating the centroid from the estimated IBM. To account for errors in mask estimation which will affect the location of the centroid of the pattern, 7 windows, centered at frames located at the centroid, centroid ±1, ±2 and ±3 frames, are selected. The CNN is then used to generate outputs for all the 7 selected windows. The output is summated to make the final classification for each pattern. This also adds to the translational invariance of the CNN. To be consistent, we use the same strategy while testing IBMs created by using the premixed signals. Note that we are not providing any duration information implicitly to the CNN, as 64 frames are longer than the longest possible utterance.

3.2 Evaluation results

Figures 3(a) and 3(b) show the recognition results using the CNN. For IBMs [Fig. 3(a)], the recognition rate is above 95% for all noise types when $\text{SNR} \geq 0$ dB. The performance is still above 90% when SNR drops to −3 dB. Only when the SNR is as low as −6 dB does the performance drop below 90%. The performance is above 85% for all noise types at all the tested SNR conditions. Note that babble noise and party noise are quite non-stationary. But our method is still able to produce good recognition results when IBMs are used, even when the SNR is as low as −6 dB.

Figure 3(b) shows the recognition results when IBMs were estimated using the method described in the previous section. For SSN, the performance ranges from 66% to 92%. The recognition rates are above 85% at $\text{SNR} \geq 0$ dB. For babble noise, the performance ranges from

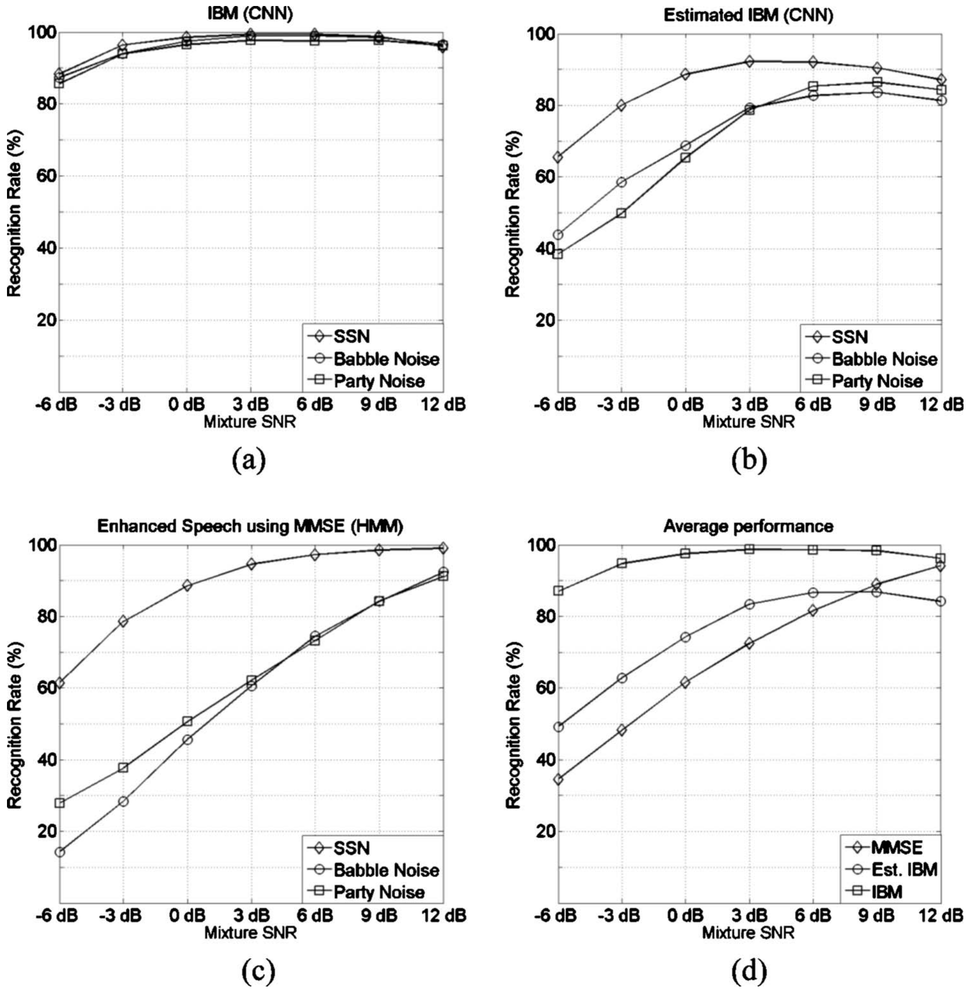

Fig. 3. Recognition results using the proposed method and a traditional HMM based method. (a) Performance of the proposed method when IBMs are used. (b) Performance of the proposed method when estimated IBMs are used. (c) Performance of the HMM based speech recognizer on enhanced speech. (d) Average performance of the proposed system and the HMM based system.

44% to 84% and for party noise, it ranges from 38% to 86%. Although the performance is expectedly not as good as the IBMs, it is still high considering the fact that the noise types are nonstationary and the SNR quite low.

To put the results in perspective, we report performance with a traditional HMM based method. Twelve word level models (1–9, zero, oh and silence) are trained using the HTK toolkit (Young *et al.*, 2009). Each model has 8 emitting states, which correspond to 10 HTK states. Output distribution in each state is modeled as a mixture of 10 Gaussians, similar to Srinivasan and Wang (2007). The only difference is that a short pause was not modeled, as the task studied in this paper is isolated, not continuous, digit recognition. The models were trained using clean speech. The feature consisted of cepstral mean normalized 12 cepstral coefficients (1–12) and normalized log energy along with their delta and acceleration coefficients (MFCC_E_D_A_Z in HTK terminology). The grammar restricted each utterance to contain just a single isolated digit. Therefore, the HMM decoder has the same knowledge about an input utterance as the CNN. The test sets consisted of mixtures at −6, −3, 0, 3, 6, 9, and 12 dB SNR, for all noise types just like those used for testing the proposed method. In each of the cases, the noisy speech was enhanced using the MMSE algorithm, which is a widely used speech enhancement algorithm (Ephraim and Malah, 1985), as our experiments showed that using such an enhancement algorithm improves recognition results in noisy conditions. Figure 3(c) shows the recognition results obtained using this MMSE-HMM method. On comparing the results with those obtained using the IBMs, we can see that it performs better only for SSN at 12 dB SNR.

When compared to the estimated IBMs, for SSN the MMSE-HMM method performs better when $\mathrm{SNR} \geq 3$ dB and for babble and party noise, it is better at 12 dB. At all other testing conditions, performance of the estimated IBMs is either comparable or better.

Figure 3(d) compares the average performance of the proposed method using IBMs and estimated IBMs, and the MMSE-HMM method. As can be seen, the best average performance is obtained using IBMs at all SNR conditions. The average performance of estimated IBMs is better till 6 dB when compared to the MMSE-HMM method. Note that MMSE algorithm works well for stationary noise (SSN). The trend would be different if the average performance was plotted only for non-stationary noise types as the MMSE algorithm has difficulty dealing with such noise types [see Fig. 3(c)].

The results suggest that the proposed system can provide a viable alternative for ASR at low SNR conditions, especially for the more challenging and realistic non-stationary noise types.

4. Concluding remarks

We have proposed a new approach to robust speech recognition. The proposed method has produced promising results for the small vocabulary task studied in this paper. Our study shows that binary patterns carry important information about the underlying phonetic content useful for ASR.

This initial study only explored isolated digit recognition, which is a relatively simple speech recognition task. Scalability to large vocabulary and continuous speech recognition tasks clearly needs to be investigated in future research. One could imagine using space displacement neural networks (LeCun *et al.*, 1998), which have been successfully used for cursive handwritten digit recognition, or using a tandem architecture (Hermansky *et al.*, 2000) to extend the proposed method to perform continuous speech recognition.

Nonetheless, the level of performance obtained in our study by using binary patterns alone, devoid of all detailed speech information, is surprising. Improvements in IBM estimation would certainly boost the performance of the proposed method. Even if binary patterns alone are proven not to be sufficient for ASR, they may well provide a complementary dimension to the traditional ASR framework in the pursuit of robust recognition.

Acknowledgments

The research described in this paper was supported in part by an AFOSR grant (FA9550-08-1-0155) and an NSF grant (IIS-0534707). We acknowledge a similar independent work described by Karadogan *et al.* (2009), which we became aware of after our model had been developed.

References and links

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT, Cambridge, MA).

Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (**2001**). “Robust automatic speech recognition with missing and unreliable acoustic data,” Speech Commun. **34**, 267–285.

Ephraim, Y., and Malah, D. (**1985**). “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” IEEE Trans. Acoust., Speech, Signal Process. **33**, 443–445.

Hermansky, H., Ellis, D., and Sharma, S. (**2000**). “Tandem connectionist feature extraction for conventional HMM systems,” in Proceedings of ICASSP, pp. 1635–1638.

Hu, K., and Wang, D. L. (**2009**). “Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction,” Technical Report No. TR51, Department of Computer Science and Engineering, The Ohio State University, Columbus, OH (available online: ftp://ftp.cse.ohio-state.edu/pub/tech-report/2009/TR51.pdf).

Karadogan, S. G., Larsen, J., Pedersen, M. S., and Boldt, J. B. (**2009**). “Robust isolated speech recognition using ideal binary masks,” Technical Report No. 5780, Department of Informatics and Mathematical Modelling, Technical University of Denmark, Kgs. Lyngby, Denmark; available at http://isp.imm.dtu.dk/staff/jlarsen/pubs/frame.htm (Last viewed 10/11/2010).

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (**1998**). “Gradient-based learning applied to document recognition,” Proc. IEEE **86**, 2278–2324.

Leonard, R. G. (**1984**). “A database for speaker-independent digit recognition,” in Proceedings of ICASSP, pp. 111–114.

Simard, P. Y., Steinkraus, D., and Platt, J. C. (**2003**). “Best practices for convolutional neural networks applied to visual document analysis,” in Proceedings of ICDAR, pp. 958–963.

Srinivasan, S., and Wang, D. L. (**2007**). “Transforming binary uncertainties for robust speech recognition,” IEEE Trans. Audio, Speech, Lang. Process. **15**, 2130–2140.

Wang, D. L., and Brown, G. J. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, edited by D. L. Wang and G. J. Brown (Wiley/IEEE, Hoboken, NJ).

Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2008**). “Speech perception of noise with binary gains,” J. Acoust. Soc. Am. **124**, 2303–2307.

Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2009**). “Speech intelligibility in background noise with ideal binary time-frequency masking,” J. Acoust. Soc. Am. **125**, 2336–2347.

Young, S., Kershaw, D., Odell, J., Valtchev, V., and Woodland, P. (**2009**). *The HTK Book (for HTK Version 3.4)* (Microsoft Corp., Redmond, WA).