Speaker Adapted Codebooks for Speech Enhancement

Chidambar B
Department of Mathematical and Computational Sciences
Sri Sathya Sai University for Human Excellence
Karnataka, India
chidambar.b@sssuhe.ac.in

D Hanumanth Rao Naidu
Department of Mathematical and Computational Sciences
Sri Sathya Sai University for Human Excellence
Karnataka, India
hanumanth@sssuhe.ac.in

***Abstract*—Speech enhancement methods employing *a priori* information of speech and noise as trained codebooks of speech and noise spectral shapes parametrized by, e.g., linear predictive (LP) coefficients have shown to perform well in non-stationary noise conditions even in single channel mode. Generally, speaker independent (SI) codebooks are employed but for applications such as mobile communication, speaker dependent (SD) codebooks are more effective. However, large amount of training data required for generating such SD models is not available in practical application. One way to overcome this limitation is to adapt available SI model to a specific speaker data using smaller amounts of training data incrementally as and when available. In this paper, we investigate the adaption of SI codebook of spectral representation of speech data to a specific target speaker using Vector Quantization Maximum *a posteriori* (VQ-MAP) algorithm and study its effect on speech enhancement performance. The experimental results indicate that VQ-MAP leads to adapted codebooks which are closer representation of a speaker than SI codebooks and enable better speech enhancement compared to SI models in codebook-based speech enhancement technique.**

I. Introduction

Speech enhancement, or noise reduction, refers to removal of noise from noisy speech. This has application in several areas such as mobile communication, hearing aids and speech recognition. Several speech enhancement techniques have been developed in the last few decades [1, 2]. Among the data driven based approaches, codebook-based speech enhancement (CBSE) techniques [3] which use prior knowledge about speech and noise power spectral densities in the form of linear predictive coefficients (LPC) have shown to provide satisfactory speech enhancement even in non-stationary noise conditions. Codebook-based approaches for speech enhancement are particularly relevant in the field of mobile communication where codebooks continue to play significant role [4]. While [3] used speaker independent (SI) codebooks, [5] introduced usage of speaker dependent (SD) models in the CBSE framework which led to better speech enhancement compared to SI models. Again, such an approach is appealing in applications like mobile communication where mostly a single speaker is involved with the device. However, generating SD codebooks requires large amount of target speaker data which is not always available in a practical scenario. This limitation can be mitigated by adapting SI models using smaller amounts of target speaker data to generate speaker adapted (SA) models.

Speaker adaptation of speech data models has been a topic of research for several decades with mainly speaker verification, speech and speaker recognition as the fields of application. The initial methods of speaker adaption dealt with Vector Quantization (VQ) models [6], however, later the focus shifted to GMM-MAP techniques [7]. Recently, Deep Neural Networks (DNN) based accoustic models have gained significance for speaker adaptation [8, 9]. In this work, we utilize codebooks, which have not been explored in the field of speaker adaptation, and apply Vector Quantization Maximum *a posteriori* (VQ-MAP) algorithm of [10] to adapt SI codebooks of LP coefficients of speech data using small amounts of speaker data. Considering that SI is trained using large amount of speaker data it is expected that the adaption will benefit from the robust structure of SI codebooks and require only small amount of speaker specific data for achieving effective adaption. In situations where the observed and modeled speakers mismatch, the framework of the CBSE [11] ensures the minimum performance upto the level of SI codebook.

VQ-MAP method uses maximum *a posteriori* approach to adapt centroids of spectral representation vectors of speaker data. It is a special case of GMM-MAP approach involving construction of a Universal Background Model (UBM) [7]. UBM represents collective spectral space of all speakers and in the case of CBSE framework relates to SI model. [10] uses Mel Frequency Cepstral Coefficients (MFCC) and applies the VQ-MAP for speaker verification. In this work, we employ VQ-MAP approach for adaptation of LPC vector centroids of speaker data and use the adapted codebooks for single channel speech enhancement. As far as we know, the speaker adapted models have not been used so far in the speech enhancement applications. The experimental results show that VQ-MAP is effective in generating speaker adapted codebooks which are better representation of a speaker's speech data compared to SI models and result in improvement of speech enhancement under the CBSE framework.

The remainder of the paper is organized as follows. The related background work is described in Section II. A brief outline of the CBSE framework and VQ-MAP algorithm is provided in Section II-A and Section II-B, respectively. In Section III, the details and results of the experiments performed for generating SA codebooks using VQ-MAP algorithm and their performance in speech enhancement are presented. Finally, Section IV provides the conclusions.

979-8-3503-3959-8/23/$31.00 ©2023 IEEE

II. Related Background Work

A. Codebook-Based Speech Enhancement

Consider an additive noise model, where the noisy signal $y(n)$ can be written as

$$y(n) = x(n) + w(n) \tag{1}$$

where $y(n)$, $x(n)$ and $w(n)$ represent the noisy speech, clean speech and noise signal, respectively, and n is the time index.

In terms of power spectral density (PSD) in the frequency (ω) domain, the above can be expressed as,

$$P_y(\omega) = P_x(\omega) + P_w(\omega) \tag{2}$$

In the CBSE method [3], the speech and noise PSDs are parametrized as LPC and stored as trained codebooks. The relation between LPC vector $\theta_x = (a_{x_0}, \ldots, a_{x_p})$, where $a_{x_0} = 1$ and p is the speech LPC model order, and the PSD of the underlying speech segment is given as $\bar{P}_x(\omega) = \frac{1}{|A_x(\omega)|^2}$ where $A_x(\omega) = \sum_{k=0}^{p} a_{x_k} e^{-j\omega k}$, and overbar in the notation $\bar{P}_x(\omega)$ is to denote the gain normalised PSD. Similar expression can be written for noise PSD $\bar{P}_w(\omega)$. Considering a pair of speech and noise PSDs given by i^{th} vector from the speech codebook and j^{th} vector from the noise codebook, the noisy PSD corresponding to the pair can be estimated as:

$$\hat{P}_y^{ij}(\omega) = g_x^{ij} \bar{P}_x(\omega) + g_w^{ij} \bar{P}_w(\omega) \tag{3}$$

where only the frequency independent level terms g_x^{ij} and g_w^{ij} corresponding to the speech and noise PSDs, respectively, are unknown. The goal of the CBSE algorithm is to estimate noisy PSD either by determining a single pair of codebook vectors from speech and noise codebook or a weighted sum of codebook spectra under a Bayesian framework, which minimizes the distortion between the observed noisy PSD $P_y(\omega)$ and the modelled PSD $\hat{P}_y(\omega)$. The corresponding speech and noise PSD estimates, $\hat{P}_x(\omega) = g_x \bar{P}_x(\omega)$ and $\hat{P}_w(\omega) = g_w \bar{P}_w(\omega)$ can then be used to construct a Wiener filter for enhancing noisy speech signal in the frequency domain:

$$H(\omega) = \frac{\hat{P}_x(\omega)}{\hat{P}_x(\omega) + \hat{P}_w(\omega)}. \tag{4}$$

A Bayesian framework was proposed in [11] wherein the speaker mismatch cases can be handled in a robust way employing both the SI and SD codebooks under the CBSE method for speech enhancement.

B. VQ-MAP

VQ-MAP is derived from GMM-MAP algorithm of [10]. For applying MAP procedure to VQ setup, a probabilistic model is constructed for the VQ model by specifying a Gaussian mixture likelihood that corresponds to the distortion criterion used in VQ model construction. The prior parameters of the Gaussian mixture as required for GMM-MAP are selected from a Universal Background Model (UBM) trained by applying a clustering algorithm such as k-means on a large number of training utterances from a number of speakers. A new speaker model is derived by adapting the well-trained UBM with that of speaker's training utterances using MAP procedure as described below.

Let UBM be represented by a set of K centroids as $U = \{u_1, u_2, \ldots, u_K\}$. Given the training data for a new speaker, $X = \{x_1, x_2, ..., x_N\}$, the adaptation is performed using the following equations in an iterative fashion,

$$q_n = \arg \min_{1 \le k \le K} ||x_n - u_k||^2 \quad 1 \le n \le N \tag{5}$$

$$\bar{x}_k = \frac{1}{|S_k|} \sum_{x_n \in S_k} x_n \tag{6}$$

$$c_k = w_k \bar{x}_k + (1 - w_k) u_k \tag{7}$$

$$w_k = \frac{|S_k|}{|S_k| + r} \tag{8}$$

where,

x_n is the n^{th} vector in X,

q_n is the cluster to which x_n belongs to,

S_k is the set of vectors in the k^{th} cluster,

$|S_k|$ is the number of vectors in the S_k cluster,

$\bar{x}_k$ is the centroid of S_k,

c_k is the adapted vector corresponding to k^{th} cluster,

w_k is the weight assigned to x_k

r is a fixed constant factor called relevance factor.

The relevance factor represents the number of training vectors that need to fall in a cluster in order to provide the computed centroid of that cluster same weight as the corresponding UBM centroid for adaptation.

III. Experimental Results

In this section, we present the details & results of the experiments performed using different codebooks - SI, SD & SA, and analyse them. We investigate whether the adapted codebooks generated by applying VQ-MAP algorithm to the SI codebook provide better representation of the speaker than SI models and whether this translates into improvement in speech enhancement in the CBSE framework.

A. Codebook training

SI & SD codebooks of speech LPC vectors were trained using speech data from the Wall Street Journal (WSJ) speech database [12]. For training the speech codebook around 180 distinct utterances of duration around 3 to 5 seconds each from a total of 50 speakers, 25 male and 25 female, were used. The SD codebook was trained using utterances from a single male speaker. The speech content used in the training of both the SI and SD codebooks was identical, and deferred only in the number of people uttering the sentences. The LP coefficients were extracted from Hann windowed segments of length 256 samples, with 50% overlap at a sampling frequency of 8 KHz. The SI and SD codebooks were trained using the Linde-Buzo-Gray (LBG) algorithm [13] with the root mean

TABLE I: Average log-spectral distortion (dB) quantization for SA codebooks trained using different relevance factors

Relevance factor	SI	SA-5-1	SA-5-2	SA-10-1	SA-10-2	SD
r=16	2.53	2.23	2.22	2.22	2.21	1.95
r=64	2.53	2.24	2.21	2.21	2.19	1.95

squared log-spectral distortion (LSD) measure as the error criterion. Speaker adapted codebooks were trained using VQ-MAP algorithm specified in Section II-B.

The adaption of SI to SA codebook was done incrementally with 5 training utterances of around 3 to 5 seconds each at a time, which adds up to around 20 seconds of adaptation data from the selected male speaker. As previously mentioned, relevance factor determines the weight assigned to adaptation data centroid as against SI centroids. A lower value of relevance factor would result in higher weightage to adaptation data which may result in over distortion of the spectral distribution of SI spectral space. Similarly, a higher relevance factor may give lesser than required weightage to the adaptation data which can slow down the adaptation. In [10], the relevance factor was fixed to 16 empirically. In our experiments, it was observed that for the amount of adaptation data used, on an average around 50 to 60 LPC vectors would fall into 256 clusters of SI codebook. Thus, using r=16 can cause over adaptation. Accordingly, relevance factor was fixed to 64. That this value doesn't slow down the adaptation was tested by comparing the LSD quantization results for 10 test utterances from the same speaker with SA trained using relevance factor values of 16 and 64. As shown in Table I, r=64 doesn't slow down the adaptation rate significantly in comparison to r=16. Here, results are shown for 4 different SA codebooks (identified with prefix SA) whose description is given below. In the rest of the sections the results are presented using relevance factor of 64.

Another factor to consider in VQ-MAP application is number of iterations to be performed for each adaptation data. In Table 1, results are presented for adaption done for upto 2 iterations using 5 and 10 training utterances and the corresponding SA codebooks are denoted as SA-5 and SA-10, respectively. SA-10 was trained by treating SA-5 as the initial SI codebook and using a set of 5 training utterances different from the ones used in training SA-5. Further, notations SA-5-1 and SA-5-2 are used to represent number of iterations to 1 and 2, respectively. Similar notation is being used for the second case of SA-10 trained using 10 adaptation utterances. In Table I, we observe that the average LSD values for first and second iterations do not differ significantly for both the cases of SA-5 as well as SA-10. So, in the rest of the sections the adaptation was done using only one iteration of VQ-MAP algorithm.

B. LPC quantization analysis for SA codebooks

Before investigating the benefit of speaker adapted codebooks in speech enhancement, we investigate whether the spectral representation of a speaker's speech data improves with the adaptation of SI codebook using training data corresponding to that speaker. To study this we compute the LPC quantization error for the test utterances when quantized using different codebooks. For this we extracted LPC vectors from set of 4 test sets - T1, T2, T3 and T4, each containing 10 distinct clean utterances with no overlap among the 4 sets, and spoken by the same speaker as the one used in training of the SD codebook. The duration of the individual utterances was between 3 to 5 seconds. LPC analysis was performed for each speech segment of around 32 ms and the resulting LPC vectors were quantized using SI, SD and SA codebooks with LSD as the error metric. The average LSD between the original and quantized LPC vectors for the 4 test data sets is shown in Figure 1 for SI, SA-5, SA-10 and SD codebooks. As can be observed, for all the 4 test data sets, the distortion value reduces with more adaptation and shifts towards distortion value corresponding to SD. This assures that the spectral representation of speech data of a speaker improves with VQ-MAP adaptation of the SI model.

In order to observe the effect of adaptation for a longer time, we performed incremental adaptation with multiple sets of 5 training utterances. We considered a total of 33 sets of such training utterances to generate SA-5, SA-10 upto SA-165 incrementally and in a sequential manner. The result for one test set T1, as shown in Figure 2, indicates that the adaptation positively takes the SI codebook spectral distribution to that of SD codebook. Also, it was observed that significant drop in LSD values occurred with the initial set of training data itself.

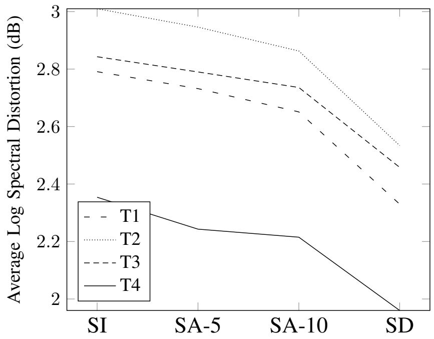

Fig. 1: Average log-spectral distortion (dB) for SA codebooks adapted using different training sets

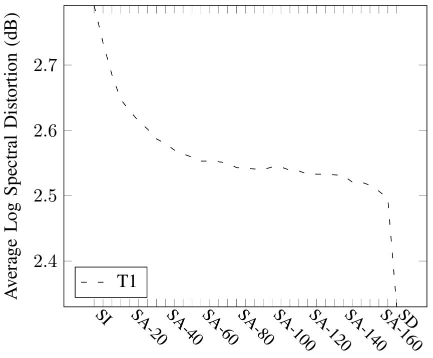

Fig. 2: Average log-spectral distortion (dB) for SA codebooks adapted incrementally and sequentially using 33 training sets containing 5 utterances each

C. Speech Enhancement using SA codebooks

In this section, we investigate whether the advantage of SA codebooks over SI translates in better speech enhancement.

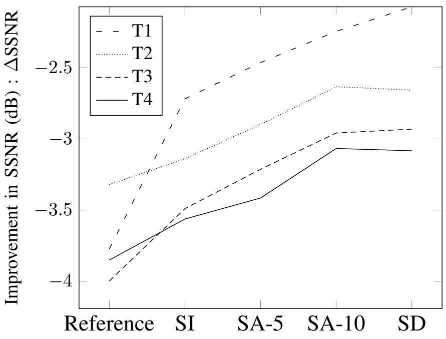

Fig. 3: ΔSSNR for SA codebooks adapted using different training sets, at 5 dB input SNR

The input noisy files for the experiments were obtained by adding non-stationary traffic noise at an SNR level of 0 and 5 dB to each of the four test sets of 10 clean utterances - T1, T2, T3 and T4, used in the previous subsection. A matching noise codebook was generated using the traffic noise data with LPC order 6 as in [11]. The noisy files were processed using SI, SA-5, SA-10 and SD codebooks using CBSE algorithm of [3]. For reference, speech enhancement reults were also obtained using a noise estimation scheme [14]. For measuring enhancement performance, two metrics were used: the improvement in segmental SNR (SSNR), referred to as ΔSSNR (in dB) and improvement in PESQ [15] measure, referred to as ΔPESQ, both averaged over all the 10 enhanced utterances of each of the four sets of test utterances. PESQ has been found to correlate well with subjective quality of the speech whereas SSNR provides objective measure for evaluating intelligibility of the speech.

Figures 3 and 4 provide ΔSSNR (in dB) and ΔPESQ values, respectively, for SA codebooks trained using 5 and 10 training utterances represented by SA-5 and SA-10 for the case of 5 dB input SNR. In both the evaluation measures, we observe CBSE based approaches provide better results than reference speech enhancement algorithm. Further for both the measures there is improvement in enhancement with adaptation of SI codebook compared to that of SI performance. This was observed when moving from SI to SA-5 as well as from SA-5 to SA-10. In the case of PESQ score, the overall improvement due to adaptation is around 0.02. The improvement is more pronounced in the case of ΔSSNR where the performance of SA-10 improves by almost 0.5 dB reaching close to that of the corresponding values of SD codebook for all the test utterances except one. Thus, a small amount of adaptation data suffices to reach close to SD performance in terms of improvement in intelligibility of the speech. Even in the case of ΔPESQ, SA-10 moves noticeably close to that of SD performance in comparison to SI in two of the four test sets. Similar pattern of observations as above were found in the case of 0 dB input SNR as shown in Tables II and III for ΔSSNR and ΔPESQ, respectively.

IV. Conclusion

Speech enhancement using model based approaches such as codebook based techniques rely on speaker independent (SI) models. However, for applications such as mobile telephony, exploiting a specific speaker's data can result in better enhancement. The practical difficulty here lies in unavailability of sufficient speaker data for generating speaker dependent (SD) models. In this work, we investigated adaptation of SI codebook to SA codebooks and using them for speech enhancement. The usage of speaker adapted codebooks in speech enhancement has not been studied so far. For performing adaption of SI codebook of linear predictive coefficients of speech data, vector quantization maximum *a posteriori* (VQ-MAP) algorithm was used. It was found that the adaptation performed provides better representation of spectral space of the speaker's speech data compared to SI codebook and also translates into improvement in speech enhancement in comparison to SI codebook. Further, it was also found that with significantly lesser amount of adaptation data compared to that required for training a SD model, the speaker adapted codebook speech enhancement performance reaches closer to that of SD codebook. This is mainly because of proper exploitation by the speaker adapted codebook of the underlying spectral space distribution of the SI models trained using large amount of speech data from many speakers. The above results clubbed with the context dependent Bayesian framework of [11] indicate that speaker adaptation is a promising approach for speech enhancement applications involving one or few speakers such as mobile phone usage. In future, we intend to extend this work to DNN-based acoustic models.

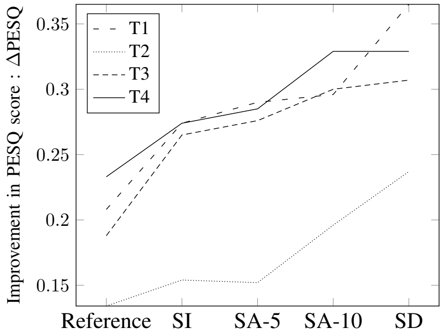

Fig. 4: ΔPESQ for SA codebooks adapted using different training sets, at 5 dB input SNR

TABLE II: ΔSSNR for SA codebooks adapted using different training sets, at 0 dB input SNR

Test sets	Reference	SI	SA-5	SA-10	SD
T1	-0.97	0.33	0.44	0.65	0.89
T2	-2.95	-2.87	-1.57	-1.45	-1.37
T3	-0.86	-0.28	-0.01	0.23	0.29
T4	-0.58	-0.41	-0.22	-0.92	0.12

TABLE III: ΔPESQ for SA codebooks adapted using different training sets, at 0 dB input SNR

Test sets	Reference	SI	SA-5	SA-10	SD
T1	0.09	0.33	0.34	0.34	0.43
T2	0.12	0.25	0.29	0.30	0.39
T3	0.23	0.33	0.36	0.36	0.39
T4	0.23	0.35	0.37	0.38	0.41

References

[1] P. Loizou, *Speech Enhancement: Theory and Practice 2nd Ed.* CRC Press, 2013.

[2] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "A comparative study of time and frequency domain approaches to deep learning based speech enhancement," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[3] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Transactions on Speech Audio Processing*, vol. 15, no. 2, pp. 441–452, 2007.

[4] S. Mabrouki, I. Dayoub, Q. Li, and M. Berbineau, "Codebook designs for millimeter-wave communication systems in both low- and high-mobility: Achievements and challenges," *IEEE Access*, vol. 10, pp. 25 786–25 810, 2022.

[5] D. Hanumantha Rao Naidu, G. V. Prabhakara Rao, and S. Srinivasan, "Speech enhancement using speaker dependent codebooks," in *2011 17th International Conference on Digital Signal Processing (DSP)*, 2011, pp. 1–6.

[6] S. Furui, "Vector-quantization-based speech recognition and speaker recognition techniques," *[1991] Conference Record of the Twenty-Fifth Asilomar Conference on Signals, Systems & Computers*, pp. 954–958 vol.2, 1991.

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[8] D. Yu and J. Li, "Recent progresses in deep learning based acoustic models," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 3, pp. 396–409, 2017.

[9] J. Deng, X. Xie, T. Wang, M. Cui, B. Xue, Z. Jin, G. Li, S. Hu, and X. Liu, "Confidence score based speaker adaptation of conformer speech recognition systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1175–1190, 2023.

[10] V. Hautamaki, T. Karkkainen, I. Krkkinen, J. Saastamoinen, M. Tuononen, and P. Franti, "Maximum a posteriori adaptation of the centroid model for speaker verification," *IEEE Signal Processing Letters*, vol. 15, pp. 162–165, 2008.

[11] H. R. D. Naidu and S. Srinivasan, "Robust Bayesian estimation for context-based speech enhancement," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, p. 35, 2014.

[12] "CSR-II (WSJ1) Complete," *Linguistic Data Consortium*, Philadelphia, 1994.

[13] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[14] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, vol. 48, no. 2, pp. 220–231, 2006.

[15] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," *International Conference in Acoustics, Speech and Signal Processing*, vol. 2, pp. 749–752, 2001.