CONTEXT-DEPENDENT QUANTIZATION FOR DISTRIBUTED AND/OR ROBUST SPEECH RECOGNITION

Chia-yu Wan, Yi Chen, and Lin-shan Lee

Graduate Institute of Communication Engineering, National Taiwan University,
Taiwan, Republic of China
chiayui@speech.ee.ntu.edu.tw, chenyi@speech.ee.ntu.edu.tw, lslee@gate.sinica.edu.tw

ABSTRACT

It is well-known that the high correlation existing in speech signals is very helpful in various speech processing applications. In this paper, we propose a new concept of context-dependent quantization, in which the representative parameter (whether a scalar or a vector) for a quantization partition cell is not fixed, but depends on the signal context on both sides, and the signal context dependencies can be trained with a clean speech corpus or estimated from a noisy speech corpus. This results in a much finer quantization based on local signal characteristics, without using any extra bit rate. This approach is equally applicable to all (scalar or vector) quantization approaches, and can be used either for signal compression in distributed speech recognition (DSR) or for feature transformation in robust speech recognition. In the latter case, each feature parameter is simply transformed into its representative parameter after quantization. In preliminary experiments with AURORA 2 and simulated GPRS channels, this concept is integrated with a recently proposed Histogram-based Quantization (HQ), the partition cells of which are also dynamic depending on local signal statistics. Significant performance improvements were obtained with the presence of both environmental noise and transmission errors.

Index Terms— Speech recognition, quantization, robustness

1. INTRODUCTION

The client-server framework for Distributed Speech Recognition (DSR) has been widely considered. In this framework, speech features are extracted and quantized in the hand-held clients, and recognition is performed at the server [1]. Robust speech recognition has also been an important topic considered by many research groups, in which the speech signals to be recognized may be seriously corrupted by additive or convolutional environmental noise. When considering the characteristics of speech signals, it is a well-known fact that the high correlation existing in speech signals is very helpful in various speech processing applications. It is also well-known that for human perception, speech is recognized based on not only the present signal values, but also on the changes in context [2]. Transform coding and differential encoding take context into consideration when performing quantization, and have been widely used for decades [2, 3, 4]. These approaches exploit inter-frame or intra-frame correlations among feature vectors and have been shown to reduce transmission rates significantly. These facts indicated that quantization approaches not using context information are relatively inadequate, because in such approaches, feature parameters with different context are quantized or transformed to the same representative value as long as they are in the same partition cell; thus signal information is not fully utilized. Therefore, properly utilizing context information in quantization to improve robustness against transmission errors and environmental noise is an important issue.

In this paper, we propose a new concept of context-dependent quantization, in which the representative parameters for each partition cell are not fixed, but are dependent on the context codewords. The context dependency of speech signals used in such quantization approaches can be trained with a clean speech corpus or estimated from a noisy speech corpus. This concept is integrated with the recently-proposed Histogram-based Quantization (HQ) [5, 6, 7], and the significantly improved performance indicates that the context-dependent quantization is very effective for both distributed and robust speech recognition.

This work is supported by the National Taiwan University Advanced Speech Technology Scholarship.

2. PROPOSED APPROACH

2.1. Context-dependent Quantization

In conventional (scalar or vector) quantization, a parameter y_t at time t (either a scalar or a vector) is mapped to a representative parameter z_i (either a scalar or a vector), which is in turn represented by a codeword or bit pattern w_t, if y_t is within a certain partition cell Q_i,

$$y_t \rightarrow Q(y_t) = z_i, w_t = b(Q(y_t)) = b(z_i), \text{if } y_t \in Q_i, \quad (1)$$

where $Q(\cdot)$ is the quantization process and $b(\cdot)$ represents the index of codeword or bit pattern. The concept of context-dependent quantization is very simple. It keeps all the original partition cells unchanged, except now the representative parameters z_i are not fixed, but are dependent on the left and right context. Assume in addition the parameter y_t has a left context parameter y_{t-1} with codeword m and a right context parameter y_{t+1} with codeword n, $y_{t-1} \rightarrow Q(y_{t-1})$, $b(Q(y_{t-1})) = m$, $y_{t+1} \rightarrow Q(y_{t+1})$, $b(Q(y_{t+1})) = n$. The representative parameter for the middle frame y_t in the partition cell Q_i is then the average of all such parameters y_t within the partition Q_i with the left and right context m and n respectively,

$$z_i^{mn} = \frac{1}{L_i^{mn}} \sum_{\substack{y_t \in Q_i \\ b(Q(y_{t-1}))=m \\ b(Q(y_{t+1}))=n}} y_t, \quad (2)$$

which is dependent on the context m and n, where L_i^{mn} is the total number of such parameters y_t in the training set. Thus z_i^{mn} is the average of the parameters with the same context codewords. This representative parameter z_i^{mn} can be trained with a clean speech corpus. In this way, context dependency among speech signals is

automatically included in the quantization process. Note that assuming there are N partition cells, for each partition cell there are now N^2 different representative parameters because there are N^2 context conditions ($m, n \in \{1, 2, \ldots, N\}$). Therefore using the left and right contexts allow for much finer representation of the parameters, although the number of bits needed remains the same. Also, the computational complexity and memory requirement on the client side are the same as those for conventional quantization because the number of partition cells is still N. This is shown in Fig. 1, in which a partition cell has many representative parameters $z_i^{m,n}$ for different contexts m and n, as compared to conventional quantization, in which a partition cell has only a single representative parameter z_i. Also, in this scheme for a received codeword sequence, every codeword is decoded considering its context codeword on both sides, and there is no problem regarding the order of decoding. For example, for the received codeword sequence, $\{w_1, w_2, w_3, \ldots\}$, $w_1w_2w_3$ are used to decode w_2, $w_2w_3w_4$ are used to decode w_3, and so on.

The above context-dependent quantization can actually be extended to decode speech signals corrupted by noise as well. Assume a noisy speech codeword sequence $[b(Q(y_{t-1})) = m, b(Q(y_t)) = i, b(Q(y_{t+1})) = n]$ is observed, where y_{t-1}, y_t, y_{t+1} are all noisy parameters, and assume that the correct codeword for the corresponding clean speech parameter $\hat{y}_t$ in the middle is $b(Q(\hat{y}_t)) = k$, where $\hat{y}_t$ is the clean speech version of y_t, and the N possible values of the codeword k has a distribution $\{P_i^{mn}(k), k = 1, 2, \ldots, N\}$. In other words, $P_i^{mn}(k)$ is the probability of the correct codeword being k (that is, $b(Q(\hat{y}_t)) = k$) when the observed noisy speech codeword sequence is $[b(Q(y_{t-1})) = m, b(Q(y_t)) = i, b(Q(y_{t+1})) = n]$. These probabilities $\{P_i^{mn}(k), k = 1, 2, \ldots, N\}$ can be easily estimated based on the frequency counts of such codeword sequences $[b(Q(y_{t-1})) = m, b(Q(y_t)) = i, b(Q(y_{t+1})) = n]$ and $[b(Q(y_{t-1})) = m, b(Q(\hat{y}_t)) = k, b(Q(y_{t+1})) = n]$ in a corpus including corresponding noisy and clean speech for some noisy conditions. With these probabilities, minimum mean squared error (MMSE) estimation for the codewords for clean feature parameters can be obtained as the conditional expectation values,

$$\begin{aligned}\hat{z}_i^{mn} &= E[z_i^{mn} \mid b(Q(y_{t-1}))=m, b(Q(y_t))=i, b(Q(y_{t+1}))=n] \\ &= \sum_k P_i^{mn}(k) z_k^{mn}, \end{aligned} \tag{3}$$

where z_i^{mn} is the context-dependent representative parameter obtained in Eq. 2, and $\hat{z}_i^{mn}$ is the MMSE estimate of the representative parameter from noisy codewords considering context dependency.

Note that the above formulation is for quantization under the DSR framework, but it applies equally to feature transformation for robust speech recognition apart from DSR, in which each original feature parameter y_t is transformed into z_i^{mn} for recognition purposes based on the quantization and its context.

All the above applies equally to all different quantization schemes. Below we apply it to the recently-proposed Histogram-based Quantization (HQ) [5, 6, 7]. First we briefly review the concept of HQ, and follow with context-dependent HQ.

2.2. Brief Review of Histogram-based Quantization

In Histogram-based Quantization (HQ) [5, 6, 7], the quantization of a feature parameter y_t at time t is based on the histogram of that feature parameter within a moving segment of the most recent T samples, $[y_{t-T+1}, \ldots, y_{t-1}, y_t] \triangleq Y_{t,T}$, up to the time t under consideration. As shown in Fig. 2, the values of these T parameters in $Y_{t,T}$ are sorted to produce a time-varying cumulative distribution function

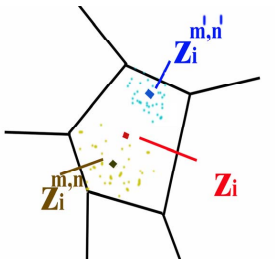

Fig. 1. Context-dependent quantization with left and right context codewords m and n.

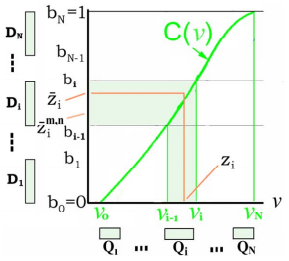

Fig. 2. Basic formulation of Histogram-based Quantization (HQ).

$C(v)$, or histogram, which changes for every time instant t. The N partition cells on the vertical scale $[0, 1]$, $\{D_i = [b_{i-1}, b_i], i = 1, 2, \ldots, N\}$ in Fig. 2, are derived from a standard Gaussian $N(0, 1)$ with cumulative distribution $C_0(v)$ via the Lloyd-Max algorithm. These partition cells on the vertical scale, $\{D_i, i = 1, 2, \ldots, N\}$, are then respectively transformed to the horizontal scale by the dynamic histogram $C(v)$ constructed with $Y_{t,T}$, to be the N partition cells $\{Q_i = [v_{i-1}, v_i], i = 1, 2, \ldots, N\}$ on the horizontal scale for the quantization of y_t, where $C(v_i) = b_i, i = 1, 2, \ldots, N$. The partition cell $Q_i = [v_{i-1}, v_i]$ on the horizontal scale is dynamic and changes for every time t via the dynamic histogram $C(v)$. In contrast, the representative parameters $\{\bar{z}_i, i = 1, 2, \ldots, N\}$ on the vertical scale $[0, 1]$ are fixed, derived from a standard Gaussian $N(0, 1)$ via the Lloyd-Max algorithm, and the representative parameters $\{z_i, i = 1, 2, \ldots, N\}$ on the horizontal scale are transformed from $\{\bar{z}_i, i = 1, 2, \ldots, N\}$ on the vertical scale by the standard Gaussian histogram $C_0(v)$; thus they are also fixed. This corresponds to the original HQ, which can also be easily extended to its vector version HVQ [5].

2.3. Context-dependent HQ

In Eq. 2 the representative parameter z_i^{mn} is determined given a set of partition cells. However, for HQ the partition cells are dynamic and varying for every time t; that is, every y_t in Eq. 2 is associated with a different set of partition cells. Fortunately, as in Fig. 2, we see that even if the partition cells Q_i for HQ are dynamic on the horizontal scale, there are another set of partition cells D_i on the vertical scale which are fixed. The dynamic histogram $C(v)$ defines the relationship between the two sets of partition cells Q_i and D_i. As a result, context-dependent HQ is easily achieved by performing Eq. 2 on the vertical scale, and then transforming it back to the horizontal scale using the standard Gaussian histogram $C_0(v)$. In other words, for context-dependent HQ we can have

$$\bar{z}_i^{mn} = \frac{1}{L_i^{mn}} \sum_{\substack{y_t \in Q_i \\ b(Q(y_{t-1}))=m \\ b(Q(y_{t+1}))=n}} C(y_t) \tag{4}$$

and

$$z_i^{mn} = C_0^{-1}(\bar{z}_i^{mn}). \tag{5}$$

Thus the contextual information represented by z_i^{mn} as obtained from Equations 4 and 5 is very similar to that of Eq. 2.

The context dependency relationships for HQ as analyzed above can then be similarly extended as in Eq. 3 to estimate the representative parameters $\hat{z}_i^{mn}$ from noisy codewords. Here, z_i^{mn} obtained from Eq. 5 can be used with the probabilities $\{P_i^{mn}(k), k = 1, 2, \ldots, N\}$ estimated from corresponding clean/noisy corpus for MMSE estimation as in Eq. 3.

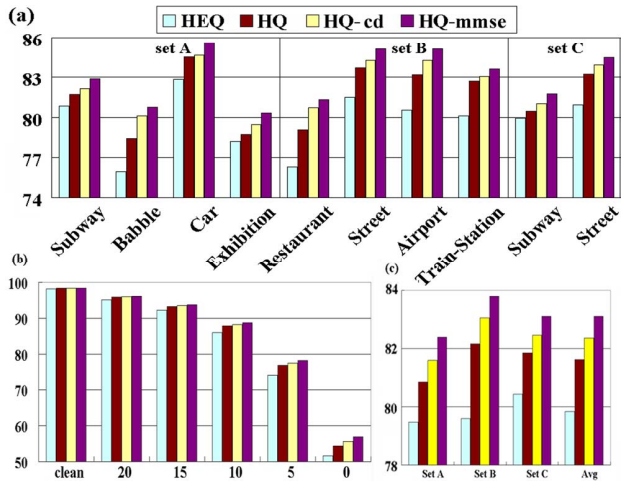

Fig. 3. Word accuracies for HEQ, HQ, HQ-cd and HQ-mmse under clean condition training: (a) averaged over all SNR values (20 dB to 0 dB) but separated for different types of noise; (b) averaged over all types of noise but separated for different SNR values; and (c) averaged over all types of noise and all SNR values (20 dB to 0 dB) for different testing sets.

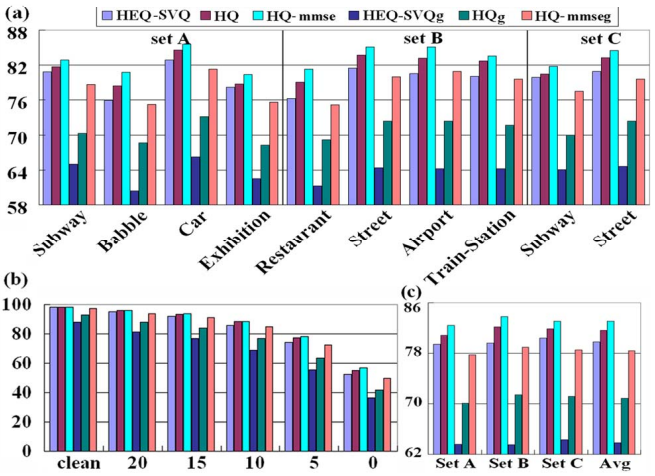

Fig. 4. Comparison of HEQ-SVQ, HQ, and HQ-mmse, and those with GPRS transmission errors (HEQ-SVQg, HQg, and HQ-mmseg): (a) averaged over all SNR values (20 dB to 0 dB) but separated for different types of noise; (b) averaged over all types of noise but separated for different SNR values; and (c) averaged over all types of noise and all SNR values (20 dB to 0 dB) for different testing sets.

SNR	Clean	20 dB	15 dB	10 dB	5 dB	0 dB
TC	98.31	95.16	89.55	70.94	43.79	18.75
HQ-mmse	98.37	96.05	93.66	88.71	78.24	56.80
TCg	93.84	84.35	73.55	52.38	27.81	9.29
HQ-mmseg	97.20	93.99	91.09	84.77	72.51	49.60

Table 1. Comparison of Transform coding (TC) and HQ-mmse, without and with GPRS transmission errors (TCg and HQ-mmseg) for different SNR values.

3. EXPERIMENTAL CONDITIONS

All experiments reported below were conducted on the AURORA 2 testing environment [8] based on a corpus of English connected digit strings. Two training conditions (clean and multi-condition) and three testing sets (sets A, B, and C) were defined in AURORA 2. Each training set consists of 8440 utterance and each testing subsets (for each noise type and each SNR condition) consists of 1001 utterance. The MFCC extraction, HMM settings and HTK-based training and testing procedures follow the Aurora 2 specifications [8]. The multi-condition training set and the corresponding clean speech training set were used to estimate the probabilities $\{P_i^{mn}(k)\}$ used in Eq. 3.

General Packet Radio Service (GPRS) was chosen as an example of wireless channels in these experiments; GPRS was developed by ETSI based on a packet switching framework to enhance the GSM system. GPRS includes four different error control coding schemes, CS1-CS4, each with a different code rate. Developed by the National Taiwan University's Wireless Communication Laboratory, the GPRS simulation software [9] used in the tests described here carefully simulated all complicated transmission phenomena, such as the propagation model, multi-path fading, Doppler spread, and so on. The experimental results presented below are based on the following simulation configurations: typical urban (TU, an environment more frequently encountered with a more severe fading problem), the client traveling at speeds of 3 km/hr, single antenna, hard decision at the receiver, and CS4 (i.e., without any protection) coding scheme, which corresponds to a transmission bit error rate of 5.3%.

4. EXPERIMENTAL RESULTS

4.1. Context-dependent HQ as a Robust Feature Transformation Method

In the first set of experiments, we considered the case of robust speech recognition apart from the DSR environment, in which context-dependent HQ was used as a feature transformation technique, that is, each feature parameter y_t, either clean or disturbed by noise, is transformed to the representative parameter z_i^{mn} in Eq. 5 or $\hat{z}_i^{mn}$ in Eq. 3, for the corresponding partition cell considering the context codewords m, n, to be used for recognition.

The results in Fig. 3 were all under clean-condition training, organized in three parts: (a) averaged over all SNR values (20 dB to 0 dB) but separated for different types of noise, (b) averaged over all types of noise but separated for different SNR values, and (c) averaged over all types of noise and all SNR values (20 dB to 0 dB) for testing sets A, B, and C, respectively. The first two bars in each set in Fig. 3 are respectively the recognition word accuracies for the well-known histogram equalization (HEQ) alone [10, 11], and the original HQ [5, 6, 7], which transforms each feature parameter y_t to the HQ representative value z_i without considering the context codewords. The next two bars are then those for context-dependent HQ, using context-dependency trained from a clean speech corpus with Eq. 5 for the third bar (HQ-cd) and using MMSE estimates trained with a multi-condition training corpus with Eq. 3 for the last bar (HQ-mmse). All the experiments reported here for HQ were based on order-statistics over segments of the most recent parameter values as mentioned in section 2.2, so there was no time delay. Although better results were obtainable if the no-delay condition was removed, they are not shown here due to space limitations. Here HEQ was performed in exactly the same way as HQ, based on a moving segment of the most recent T parameters, and the same value of $T = 100$ was used.

It can be found that HQ (2nd bar) consistently outperformed HEQ (1st bar) (this was verified earlier [6]), while context-dependent HQ (both HQ-cd and HQ-mmse in the 3rd and 4th bars) consistently and significantly outperformed HEQ: in particular MMSE estimation

trained with a noisy corpus (4th bar) resulted in much more robust features for recognition. Increasing improvements are apparent in Fig. 3 in all cases. In addition, context-dependent HQ trained with clean speech (HQ-cd, 3rd bar) offered greater improvement than original HQ (HQ, 2nd bar) for speech-like noise such as babble, restaurant, and airport, probably because the context-dependent characteristics for these types of noise have been more or less included in the transformation. Furthermore, HQ-mmse (4th bar) consistently outperforms HQ-cd (3rd bar) (Fig. 3(c)), which verifies that the context dependency trained from noisy corpora is useful even for unseen noisy environments (e.g. sets B and C).

4.2. Context-dependent HQ as a Feature Quantization Method for DSR

We next considered context-dependent HQ as a feature quantization method in DSR. In Fig. 4 in each set the first three bars are respectively the word accuracies for the well-known HEQ followed by the conventional SVQ (HEQ-SVQ), original HQ (the same as the 2nd bar in Fig. 3), and context-dependent HQ with MMSE estimation (HQ-mmse, the same as the 4th bar in Fig. 3), all at 4.4 kbps without transmission errors, and the next three bars (HEQ-SVQg, HQ-g, HQ-mmseg: the label "g" indicates GPRS) are those suffering from GPRS transmission errors for a client traveling at 3 km/hr. Fig. 4 (a) is averaged over all SNR values (20 dB to 0 dB) but separated for different types of noise, (b) is averaged over all types of noise but separated for different SNR values, and (c) is averaged over all types of noise and all SNR values (20 dB to 0 dB) for testing sets A, B, and C, respectively.

We first examined the effect of quantization and compression on recognition accuracy, assuming there were no transmission errors. The performance of original HQ (2nd bar) consistently outperformed HEQ-SVQ (this was also shown previously [6]), while HQ-mmse (3rd bar) was consistently and significantly better than original HQ, as shown in Fig. 4(a)-(c). This verifies the effectiveness of context-dependency. Improvements were even more significant for lower SNR cases (Fig. 4(b)), and for several types of non-stationary noise (Fig. 4(a)), which indicates where context-dependency is more helpful. We then examined the effect of transmission errors in the last three bars in Fig. 4. For HEQ-SVQ, the performance degradation caused by GPRS (4th bar compared to 1st bar) is more serious for lower SNRs. Clearly, features corrupted by noise are more susceptible to transmission errors. The improvements that HQ and context-dependent HQ offered over HEQ-SVQ when transmission errors were present (5th, 6th bars to 4th bar) are consistent and very significant. For example, in the case of 10 dB SNR with GPRS, HQ-mmseg (6th bar) offered an accuracy of 84.77% compared to 69.84% for HEQ-SVQg (4th bar). In addition, it is interesting that the improvements offered by HQ-mmse over HQ when transmission errors were present (6th bar to 5th bar) are much more significant as compared to those comparison without transmission errors (3rd bar to 2nd bar). This indicates that context-dependency among speech codewords is actually very strong, and remains helpful even after heavy disturbance due to environmental noise and transmission errors, and the error propagation problem is not serious here. This is probably because even if there are erroneous context codewords, they may only change the representative parameter z_i^{mn} of the current frame within the same partition cell Q_i in Fig. 1, which is actually very limited. Also, the decoding here used only local context codewords, i.e., based on the two neighboring undecoded codewords only; thus erroneous codewords actually do not propagate. It is clear from Fig. 4 that HQ-mmse is robust against both environmental noise and transmission errors.

Also shown in Table 1 are the detailed word accuracies of transform coding (TC) [4] compared with HQ-mmse, either without or with GPRS transmission errors for all SNR values, average over all noise types. The performance of TC seriously degrades when transmission errors are present (3rd row vs. 1st row), probably because exploiting speech correlation by grouping several consecutive frames into one block and quantizing them together may be sensitive to transmission errors. In contrast, error propagation is not a serious problem here for HQ-mmseg (the performance degradation is much smaller for the comparison of 4th and 2nd rows).

5. CONCLUSIONS

We have proposed context-dependent quantization, a new concept for distributed and/or robust speech recognition. Improved recognition performance was obtained consistently across a wide range of environmental noise and transmission error conditions.

6. REFERENCES

[1] V.V. Digalakis, L.G. Neumeyer, and M. Perakakis, "Quantization of cepstral parameters for speech recognition over the world wide web," *IEEE Select. Areas Commun.*, vol. 17, no. 1, pp. 82–90, January 1999.

[2] Q. Zhu and A. Alwan, "An efficient and scalable 2D-DCT based feature coding scheme for remote speech recognition," in *Proc. ICASSP*. IEEE, 2001, pp. 113–116.

[3] I. Kiss and P. Kapanen, "Robust feature vector compression algorithm for distributed speech recognition," in *Proc. Eurospeech*, 1999, pp. 2183–2186.

[4] W.-H. Hsu and L.-S. Lee, "Efficient and robust distributed speech recognition (DSR) over wireless fading channels: 2D-DCT compression, iterative bit allocation, short BCH code and interleaving," in *Proc. ICASSP*, 2004, pp. 69–72.

[5] C.-Y. Wan and L.-S. Lee, "Histogram-based quantization (HQ) for robust and scalable distributed speech recognition," in *Proc. Interspeech*, 2005, pp. 957–960.

[6] C.-Y. Wan and L.-S. Lee, "Joint uncertainty decoding (JUD) with histogram-based quantization (HQ) for robust and/or distributed speech recognition," in *Proc. ICASSP*, 2006, pp. 125–128.

[7] C.-Y. Wan, Yi Chen, and L.-S. Lee, "Three-stage error concealment for distributed speech recognition (DSR) with histogram-based quantization (HQ) under noisy environment," in *Proc. ICASSP*, 2007, pp. 877–880.

[8] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000*, Sep. 2000.

[9] J.-H. Chen, "Receiver design and simulation analysis of GPRS physical layer," *Master Thesis, National Taiwan University*, June 2001.

[10] S. Molau, M. Pitz, and H. Ney, "Histogram based normalization in the acoustic feature space," in *Proc. ASRU*, 2001.

[11] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.