CHANNEL-MAPPING FOR SPEECH CORPUS RECYCLING

Osamu Ichikawa[1], Steven J. Rennie[2], Takashi Fukuda[1], Masafumi Nishimura[1]

[1]IBM Research – Tokyo, Toyosu, 135-8511, JAPAN
[2]IBM T. J. Watson Research Center, Yorktown Heights, NY 10598
ichikaw@jp.ibm.com, sjrennie@us.ibm.com, {fukuda1, nisimura}@jp.ibm.com,

ABSTRACT

The performance of automatic speech recognition (ASR) is heavily dependent on the acoustic environment in the target domain. Large investments have focused on ways to record speech data in specific environments. In contrast, recent Internet services using hand-held devices such as smartphones have created opportunities to acquire huge amounts of "live" speech data at low cost. There are practical demands to reuse this abundant data in different acoustic environments. To transform such source data for a target domain, developers can use channel mapping and noise addition. However, channel mapping of the data is difficult without stereo mapping data or impulse response data. We tested GMM-based channel mapping with a vector Taylor series (VTS) formulation on a per-utterance basis. We found this type of channel mapping effectively simulated our target domain data.

Index Terms— Speech recognition, feature adaptation, channel normalization, noise reduction

1. INTRODUCTION

ASR products started with desk-top dictation into a telephony system and were later built into embedded systems for automobiles. This domain-focused approach has led to large investments in acquiring speech data for specific domains. The same need for large investments has also prevented ASR products from being used for more languages around the world.

Nowadays, smartphones and tablet devices are in widespread use and the situation is changing. Messaging and search applications using human speech are becoming routine. Many of them use an Internet connection to send speech data to an ASR server and receive recognized text. Such an ASR server could easily and inexpensively collect huge amounts of natural and spontaneous speech. Currently, this data is only used to retrain the acoustic model of the ASR system, but there is enormous potential in transforming the data into corpora for other application domain.

The traditional approach to synthesize the data for a target domain data is to first convolve the impulse responses and then add the environmental noise [1]. This is a straightforward approach to compensate for the channel and noise characteristics. However, it is not suitable for our purposes, because the channel characteristics of the input sources are too diverse for any single impulse response, and the input data is never completely clean. Stereo mapping can reduce these problems. For example, SPLICE [2] estimates the cepstrum bias between the source domain and the target domain, based on an a priori model trained with stereo data, at least in cases where stereo data is available.

Instead of such data transformations, we can adapt an acoustic model for a target domain using a small amount of target data. Parallel Model Combination (PMC) [3] and Vector Taylor Series (VTS) adaptation [4] can transform models for noisy environments. Maximum Likelihood Linear Regression (MLLR) [5] adjusts the model parameters to maximize the likelihood of the adaptation data. Seltzer et al. proposed cascaded MLLR (CMLLR) [6] to adapt to the environmental characteristics and speaker characteristics independently. It remains a challenging task to separate the environmental characteristics and the speaker characteristics without identifying the speakers and labeling the environments. In the field of speaker verification, Feature Mapping [7] and Joint factor analysis [8] tackled this challenge.

Unlike many of the previous projects, we are not exploring a rigorous separation of channel characteristics and speaker characteristics. We accept a target domain GMM as the a priori knowledge of the channel characteristics to mix with minimum speaker variations.

2. AUDIO MAPPING PIPELINE

Channel and noise characteristics are the major factors to map speech data into a target domain. Fig. 1 shows our mapping pipeline to work with the factors.

It is crucial to compensate the channel data before the noise. To illustrate this problem, the same automobile noise was added to two different kinds of speech data. Fig. 2 shows the distributions of the noises after cepstrum mean normalization (CMN) for each utterance. Even though the same noise was added, the resulting signals are quite different due to the run-time channel normalization (such as

CMN). This motivated us to tailor the channel characteristics first so to have similar signals at the decoder.

3. CHANNEL MAPPING

The basic idea is as follows: the source domain has multiple utterances but they are so dissimilar that they cannot be modeled with one model. There is a target domain with a little data from which one can estimate a GMM. For each utterance in the source domain, find channel bias and amplitude that can map the utterance to the target domain. To do this, use VTS [9][10] to map the target domain Gaussians to the source domain. This incorporates bias and amplitude as unknown parameters and ML is used to get the best estimate of them. Subsequently, the source utterance is transformed to the target domain by applying the channel bias and amplitude. The transformed source data is then used as training data to estimate a full-fledged acoustic model for decoding the sentence in the target domain.

3.1. Bias-only formulation

First we work with the channel bias. For the source domain, observation y is described using clean speech x, channel h, noise n, and a mismatch function G in the cepstrum domain as

$$y = h + x + G(x + h, n), \tag{1}$$

$$G(x,n) = C\log\left(1 + \exp\left(C^{-1}(n - x)\right)\right). \tag{2}$$

The matrix C is a Discrete Cosine Transform (DCT) matrix. The target domain clean speech $\hat{y}$ can be characterized as

$$\hat{y} = \hat{h} + x\,. \tag{3}$$

Using Equations (1) and (3), we have

$$y = \hat{y} + \left(h - \hat{h}\right) + G(x + h, n) = \hat{y} + c + G(\hat{y} + c, n). \tag{4}$$

The channel bias c is now defined as

$$c = h - \hat{h}\,. \tag{5}$$

We set c so as to minimize the auxiliary function Q.

$$Q = E\left[\sum_k^K \rho_k(y) \cdot \left\{\sum_d^D \left(y_d - \mu_{y,k,d}\right)^2 / \Sigma_{y,k,d} + \log\left|\Sigma_{y,k}\right|^{\frac{1}{2}}\right\}\right] \tag{6}$$

The expectation is calculated using all of the speech frames in each utterance. The posterior probability ρ_k is for the k-th Gaussian, calculated as

$$\rho_k(y) = \gamma_k \cdot \mathrm{N}\left(y; \mu_{y,k}, \Sigma_{y,k}\right) \Big/ \sum_{k'} \gamma_{k'} \cdot \mathrm{N}\left(y; \mu_{y,k'}, \Sigma_{y,k'}\right), \tag{7}$$

where γ_k is the priori probability, $\mu_{y,k,d}$ is a mean statistic for the d-th component of the k-th Gaussian in the source domain, and $\Sigma_{y,k,d}$ is the variance. We used the diagonal covariance approximation. Because the GMM is for the target domain, the source domain statistics need to be derived from the target domain statistics $\mu_{\hat{y},k,d}$ and $\Sigma_{\hat{y},k,d}$ using Eqn (4).

$$\mu_{y,k,d} \cong \mu_{\hat{y},k,d} + c_d + G\left(\mu_{\hat{y},k} + c, \mu_n\right)_d \tag{8}$$

$$\Sigma_{y,k,d} \cong \sum_l \left(\delta_{d,l} - F\left(\mu_{\hat{y},k} + c, \mu_n\right)_{d,l}\right)^2 \cdot \Sigma_{\hat{y},k,l} + \sum_l F\left(\mu_{\hat{y},k} + c, \mu_n\right)_{d,l}^{\ 2} \cdot \Sigma_{n,l} \tag{9}$$

$$F(x,n)_{i,j} = \sum_k C_{i,k} \cdot \left(C^{-1}_{k,j}\right) \cdot \frac{\exp\left(\sum_l C_{k,l}^{\ -1}\left(n_l - x_l\right)\right)}{1 + \exp\left(\sum_l C_{k,l}^{\ -1}\left(n_l - x_l\right)\right)} \tag{10}$$

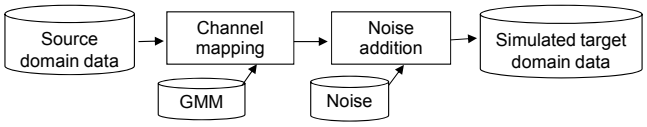
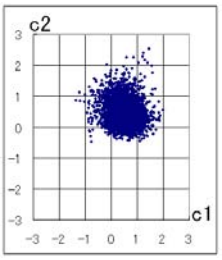
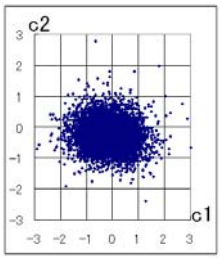


Fig. 1. Audio mapping pipeline.



(a) Automobile noise normalized by parked-car speech mean.



(b) Automobile noise normalized by handheld speech mean.

Fig. 2. Distributions of the noise after CMN, plotted only for the lower cepstrum c1 and c2.

The noise statistics μ_n and $\Sigma_{n,d}$ are calculated for the non-speech segments. We used power-based segmentation here. We iteratively estimated the channel bias c for the d-th component by differentiating Eqn (6) w.r.t. c_d and setting it to zero. The indirect derivatives of F were ignored, so the second term in Eqn (6) can also be ignored, because $\Sigma_{y,k}$ is treated as a constant. The mapped output $\tilde{y}$ is given by

$$\tilde{y} = y - c\,. \tag{11}$$

3.2. Bias and amplitude formulation

We can now introduce the amplitude into the channel mapping formulation. This is an analogy of some adaptation techniques such as Mean and Variance Normalization (MVN) [11] or a diagonal MLLR [12]. Thus, Eqn. (4) can be extended as

$$y = a * \hat{y} + c + G\left(a * \hat{y} + c, n\right), \tag{12}$$

where the $*$ denotes a component-wise product and a is the amplitude. The source domain statistics are given by

$$\mu_{y,k,d} \cong a_d \cdot \mu_{\hat{y},k,d} + c_d + G\left(a * \mu_{\hat{y},k} + c, \mu_n\right)_d, \tag{13}$$

$$\Sigma_{y,k,d} \cong \sum_l a_l^{\ 2} \cdot \left(\delta_{d,l} - F\left(a * \mu_{\hat{y},k} + c, \mu_n\right)_{d,l}\right)^2 \cdot \Sigma_{\hat{y},k,l} + \sum_l F\left(a * \mu_{\hat{y},k} + c, \mu_n\right)_{d,l}^{\ 2} \cdot \Sigma_{n,l}\,. \tag{14}$$

Similar to the bias-only case, we take the derivatives of Eqn (6) with a and c set to zero. Starting with the initial values of c=0 and a=1, the mapping parameters c and a are

updated iteratively. In our experiment, we updated *c* first, then we updated *a*. For *c*, all 13 of the components were updated. For *a*, only the 0th to 2nd components were updated, because the lower cepstra are the major and stable representations of the channel. In order to work with possible problematic utterances, we limited the range of *a* from 0.5 to 2.0.

The mapped output $\tilde{y}$ is given by

$$\tilde{y} = a^{-1} * (y - c) \tag{15}$$

3.3. Gender-dependent formulation

Similar to the gender-dependent labeling (GDL) [13] for an ASR decoder, we can prepare separate GMMs for male and female voices. They are combined with the gender weights λ_g.

$$Q = E\left[\sum_{g=f,m} \lambda_g \sum_{k}^{K} \rho_{g,k}(y) \cdot \left\{\sum_{d}^{D} \left(y_d - \mu_{y,g,k,d}\right)^2 / \Sigma_{y,g,k,d} + \log\left|\Sigma_{y,g,k}\right|^{\frac{1}{2}}\right\}\right] \tag{16}$$

The gender weights λ_g are updated based on the posterior probabilities of the GMMs during the iterations to determine the mapping parameters as

$$\lambda'_g = E\left[\sum_{k} \gamma_{g,k} \cdot \mathrm{N}\left(y; \mu_{y,g,k}, \Sigma_{y,g,k}\right)\right], \tag{17}$$

$$\lambda_g = \lambda'_g \Big/ \sum_{g'=f,m} \lambda'_{g'} \cdot \tag{18}$$

We may optionally introduce softmax for faster gender determination.

4. NOISE ADDITION

To simulate the noise characteristics, ambient noises need to be recorded in situations appropriate for the target domain. Preferably, the data should be recorded with the same cars and microphones as those used for training the GMM. Noises are randomly selected and mixed with the speech data after channel-mapping. We did not remove the noise in the source data beforehand, because it was sufficiently small. The mixing weight is adjusted for each utterance so as to conform to the overall target SNR distribution.

5. EXPERIMENTS

We performed two kinds of experiments. Our preliminary experiment used the relatively small public database CENSREC-3 [14]. Since it is a well conditioned database, we could accurately measure the effectiveness of our proposed channel-mapping methods. Our more realistic experiment used a large amount of LVCSR data. The recording environments are quite diverse, including offices, home, cars, restaurants, and trains, but there is no location information tagged to each utterance.

5.1. Preliminary experiment using CENSREC-3

We evaluated the channel-mapping part in isolation using CENSREC-3, which is a standard evaluation framework for isolated Japanese word recognition in automobiles. It has both training and testing data for automatic speech recognition using multi-style trained acoustic models.

In this experiment, we studied channel-mapping from close-talk data to far-field data. For training, a total of 3,608 utterances spoken by 293 drivers (202 males and 91females) were recorded in parked cars. The utterances recorded with the close-talk microphone were transformed by the channel-mapping methods to simulate far-field data. The GMM for the mapping has 256 Gaussians and is trained with 500 randomly selected utterances from the clean far-field microphone data. For testing, a total of 898 utterances spoken by 18 speakers (8 males and 10 females) were recorded in parked cars with the far-field microphone. This does not include noisy cases such as air-conditioner or car-audio. The recognition grammar is a list of 50 words.

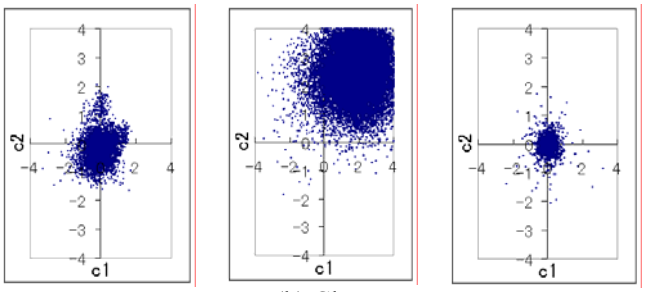
The front-end was configured with the default settings of CENSREC-3. The sampling frequency was 16 kHz and we used 39-dimension features (12 mel-cepstrum + log power, with Δ and ΔΔ), with and without subtracting the cepstrum mean.

Table 1 shows the experimental results. The R1 is the matched case using far-field data for the training. This gives us the upper limit for our trials. The R2 is our baseline using close-talk data for the training without using channel-mapping. Due to the channel mismatch, the accuracy was severely degraded, even though it was decoding fairly clean speech. It should be noted that the simple CMN was still insufficient. The R3 to R5 use our proposed channel-mapping and showed good recovery. For the R3 values, a bias-only implementation was used for the mapping. R4 improved the result with gender-dependent GMMs. This is supporting evidence that the GMM in our approach still accounts for both the speaker and channel characteristics. For the R5, both the bias and amplitude are considered in the mapping. This shows the proposed channel-mapping method works well in the combination of CMN and reduced errors by 58%.

5.2. Testing a larger volume of LVCSR data

We tested an acoustic model for automobile environments built from abundant English LVCSR data using some a priori knowledge about the channel and the noise characteristics of the target domain, which are specifically defined by a GMM and a noise recording.

We could build an acoustic model from scratch with all of the LVCSR data processed by the audio-mapping pipeline described in Section 2. However, this naïve approach with so much data would take an unacceptably long time for the repeated build processes with discriminative training. Therefore, in our experiment, we built a baseline acoustic model using all of the LVCSR data without any mappings and we adapted this baseline model



(a) Parked-car data. (b) Clean LVCSR data (SNR > 25dB). (c) Clean LVCSR data with channel-mapping.

Fig. 3. Distributions of the channel bias measured with the parked-car GMM. Only for c1 and c2 are plotted here.

with adaptation data generated from a small part of the LVCSR data by using the audio-mapping pipeline.

The baseline acoustic model was discriminatively trained with a large amount of the LVCSR data. This data represented several thousands of hours of speech. It contains roughly 400k of Gaussians with 20k of quinphone context-dependent states. The sampling frequency was 16 kHz. The front-end acoustic features are 48-dimensional MFCC including static and dynamic features.

For the adaptation data, 70 hours of utterances were randomly selected from the high SNR subset of the LVCSR data. The selection criteria of the SNR were 25 dB. The utterances were processed by channel-mapping and noise-addition.

The GMM for the channel-mapping has 256 Gaussians and was trained with a randomly selected 500 utterances (about 1 hour) from 10 hours of recordings spoken by 28 speakers (15 males and 13 females) in parked cars. For the noise sources, 15 noise files were recorded in two cars in different driving conditions. Our results are for an in-house test set for an in-car messaging task in English. A total of 8976 utterances spoken by 44 speakers (22 males and 22 females) were recorded with a far-field microphone in real moving and parked cars for this test.

The channel characteristics can be observed as a bias measured with a reference GMM. Fig. 3 shows the distributions of the bias measured by the channel-mapping. They are plotted only for c1 and c2, because the low cepstra are the essential parts of the channel. Since they were measured with the parked-car GMM, the 3(a) plots for the parked car are distributed around the original point. In contrast, the 3(b) plots for the LVCSR data are widely shifted. However in 3(c) they are successfully aligned with the original point after the channel-mapping.

Table 2 shows the word error rates (WER) when the models decode the in-car messaging data. H1 used the baseline acoustic model, whose training data includes various noisy cases. H2 to H4 used the adapted models. H2 used noise addition for the clean part of the LVCSR data to generate the adaptation data. H3 also used the channel-mapping of the bias-only implementation from Section 3.1. There was a significant improvement from H2 to H3 in favor of the channel-mapping. For H4, both bias and amplitude are considered for the mapping. For the amplitude, the 0^{th} to 2^{nd} components were shared in this case. We confirmed the amplitude implementation was better, though the improvement was small in this experiment.

6. CONCLUDING REMARKS

Although current ASR decoders are equipped with powerful channel compensation, the channel-mapping before noise-addition is a critical step for "speech corpus recycling" that reuses a huge data collection from the Internet for a different domain (such as our automotive domain). We investigated the advantages of the GMM-based channel-mapping. The GMM trained with a small amount of the target domain data is regarded as likely to include the target domain channel characteristic slightly mixed with the speaker characteristics. Unlike most of the previous research, our approach does not depend on a rigorous separation of the channel and speaker characteristics. Our experiment tested GMM-based channel-mapping for each utterance, and it showed a significant capability to simulate the target domain data. The GMM-based channel-mapping was further extended for a gender dependent model and to a "bias and amplitude" model inspired by MVN and diagonal MLLR.

7. ACKNOWLEDGEMENT

The present study was conducted using the CENSREC-3 database developed by the IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group.

Table 1. Results of decoding the clean far-field data in CENSREC-3 with acoustic models trained with far-field or close-talk data

	Training data	Channel-mapping	Accuracy %	
			no CMN	CMN
R1	Far-field	No	99.7	99.7
R2	Close-Talk	No	92.3	95.2
R3	Close-Talk	Yes, bias only	93.1	96.0
R4	Close-Talk	Yes, bias only. Gender Dependent.	94.5	96.8
R5	Close-Talk	Yes, bias and amplitude. Gender dependent.	96.4	98.0

Table 2. Results of decoding in-car messaging data with the baseline acoustic model and the adapted acoustic models

	Acoustic Model	WER %	
H1	Baseline model	17.66	
	Acoustic Model adapted by high SNR part of LVCSR data (70 hours) with :	MLLR	MLLR + MAP
H2	noise addition only	17.65	17.98
H3	channel mapping (bias) + noise	16.86	16.72
H4	channel mapping (bias and amp.) + noise	16.99	16.68

8. REFERENCES

[1] V. Stahl , A. Fischer, and R. Bippus, "Acoustic Synthesis of Training Data for Speech Recognition in Living Room Environments," Proc. of ICASSP, Vol. 1, pp. 285-288, 2001.

[2] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE Algorithm on the Aurora2 Database", Proc. of Eurospeech, pp. 217-220, 2001.

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

[4] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," Proc. of ICASSP, Vol. 2, 1996.

[5] M. J. F. Gales, "Maximum likelihood linear transformations for HMMbased speech recognition," Computer Speech and Language, Vol. 12, pp. 75–98, 1998.

[6] M. L. Seltzer and A. Acero, "Factored Adaptation for Separable Compensation of Speaker and Environmental Variability", Proc. of ASRU, pp. 146-151, 2011.

[7] D. A. Reynolds, "Channel Robust Speaker Verification via Feature Mapping," Proc. of ICASSP,vol.II, pp. 53-56, 2003.

[8] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification," Proc. of ICASSP, Vol. 1, pp. 37-40, 2004.

[9] B. Raj, E. Gouvêa, and R. M. Stern, "Cepstral compensation using statistical linearization," Proc. of the ETRW, 1997.

[10] D. Y. Kim, C. K. Un, and N. S. Kim, "Speech recognition in noisy environments using first-order vector Taylor series," Speech Communication, Vol. 24, pp. 39-49, 1998.

[11] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," Speech Communication, Vol. 25, pp.133–147, 1998.

[12] M.J.F. Gales, D. Pye and P.C. Woodland, "Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation," Proc. of ICSLP, Vol. 3, pp.1832-1835, 1996.

[13] P. Olsen and S. Dharanipragada, "An Efficient Integrated Gender Detection Scheme and Time Mediated Averaging of Gender Dependent Acoustic Models," Proc. of ICASSP, pp. 2509-2512, 2003.

[14] M. Fujimoto, et al., "CENSREC-3: Data collection for in-car speech recognition and its common evaluation framework", Proc. of RWCinME, pp. 53-60, 2005.