PHONEME CLASS BASED ADAPTATION FOR MISMATCH ACOUSTIC MODELING OF DISTANT NOISY SPEECH*

Seçkin Uluskan, John H. L. Hansen

Center for Robust Speech Systems
University of Texas at Dallas, Richardson, TX, USA
sxu092020@utdallas.edu, john.hansen@utdallas.edu

Abstract

A new adaptation strategy for distant noisy speech is created by phoneme class based approaches for context-independent acoustic models. Unlike the previous approaches such as MLLR-MAP adaptation which adapts acoustic model to the features, our phoneme-class based adaptation (PCBA) adapts the distant data features to our acoustic model which has trained on close microphone TIMIT sentences. The essence of PCBA is to create a transformation strategy which makes the distribution of phoneme-classes of distant noisy speech be similar to those of close microphone acoustic model in thirteen dimensional MFCC space (mostly in c0-c1 plane). It creates a mean, orientation and variance adaptation scheme for each phoneme class to compensate the mismatch. New adapted features, and new and improved acoustic models which are produced by PCBA are outperforming those created by MLLR-MAP adaptation for ASR and KWS. And PCBA offers a new powerful understanding in acoustic-modeling of distant speech.

Index Terms: phoneme class, distant noisy speech, mismatch acoustic modeling, feature adaptation

1. Introduction

Phoneme class based approaches have been previously used in speech studies to address different problems. It is believed that recognition of phoneme-classes provides speech processing tasks with additional acoustic information that can be utilized to improve the speech enhancement [1]. In the earliest studies, a hidden-Markov model based phoneme class detection algorithm was proposed to help for speech enhancement of noisy speech [2]. In addition to speech enhancement, phoneme-class based approaches have also been used more recently in emotion detection [3]. As a result, phoneme classes can be used in variety of speech processing studies because of the informative and productive characteristics of phoneme-class based approaches.

At this point, it would be useful to consider employing phoneme-classes to find a new mismatch compensation and acoustic modeling strategy for distant speech. It is possible to analyze the c0-c1 scatter of close talk microphone speech with a synchronized distant microphone equivalent where it is observed essentials differences between these two scatters. While the close talk speech is distributed over a large area within a specific (umbrella-like) shape in the c0-c1 plane, the distant speech is restricted in a relatively small and circular region. In order to explore the usefulness of this observation, we wanted to conduct some experiments which adapt c0-c1 coefficients of distant data to those of close talk microphone data by taking only the c0 and c1 dimensions of the MFCCs into account. Even though some improvement observed, the resulting speech recognition performance was not as effective as expected. It was reasoned that this is because the location of a frame in c0-c1 plane is mostly governed by its phoneme-class. Finally, the results from this probe investigation suggest that any study which does not take the phoneme-class information into account might not yield effective results.

Inspired by our probe analysis of the c0-c1 scatter space between distant speech and close talk microphone speech, we now turn to propose a new type of adaptation and acoustic modeling strategy for distant speech. This adaptation technique aims to make the distribution of the phoneme-classes of distant speech similar to the close talk microphone acoustic model in a thirteen dimensional MFCC space.

2. Methods:

2.1. Determination of Phoneme Classes

The major analysis here was conducted within the c0-c1 space by investigating the behavior close talk microphone acoustic model and distant data MFCCs. First by investigating the close talk microphone acoustic model, the English phonemes were divided into 5 phoneme classes based on their position in the c0-c1 plane. The context independent (CI) acoustic model which is under investigation is based on 40 phonemes (monophones) with 3 states per phoneme and one mixture per state. We define the position of a phoneme as the mean value of its second state (out of 3 states) in the related MFCC dimension. The names of the phoneme-classes are based on the dominant group of phonemes occupying each class.

Practical Phoneme Class Determination by c0-c1 plane	
1. Class: Fricatives (most of the fricatives and affricatives)	CH, F, JH, S, SH, TH, Z, ZH
2. Class: Plosives (or stops)	P, K, T, B, G, D
3. Class: Mix (nasals and some others)	DH, HH, M, N, NG, V, W
4. Class: Silence	SIL
5. Class: Vowels (and most of the glides and liquids)	AA, AE, AH, AO, AW, AY, EH, ER, EY, IH, IY, L, OW, OY, R, UH, UW, Y

Table 1. *Phoneme-Classes based on the location of phonemes in c0-c1 plane.*

After obtaining a scheme which divides all existing 40 phonemes into 5 classes with the help of an acoustic model trained using TIMIT data [4], we labeled each frame of our

* This project was funded by AFRL under contract FA8750-12-1-0188 (Approved for public release, distribution unlimited: 88ABW-2012-2076), and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

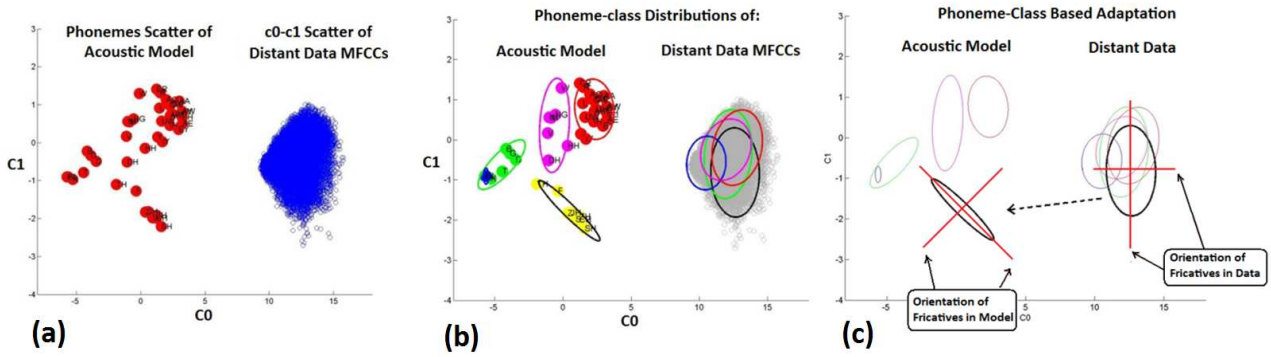

Figure 1: Illustration of Phoneme Class-Based Adaptation (PCBA) for distant speech. (a) scatter plots of phonemes of close talk microphone acoustic model in and distant speech MFCCs in c0-c1 plane; (b) distribution and alignment of different phoneme classes for both acoustic model and distant data; (c) PCBA based mean, orientation and variance adaptation for fricatives of distant data towards close talk acoustic model.

distant data with their corresponding phoneme-classes. In order to obtain very accurate phoneme-class labeling, we are determining the locations of phoneme-classes by using the synchronized close talk microphone audio streams and then transferring this labels to our distant microphone audio streams. Once we complete these labels, we can formulate "phoneme class based adaptation (PCBA)" solution for distant-based data.

2.2. Phoneme-Class based Adaptation (PCBA)

We employ Fig. 1 to illustrate the overall procedure for PCBA of distant speech. Consider Fig.1.a, where it is observed that a significant difference between the close talk microphone acoustic model and the distant captured data in the c0-c1 plane. Here it is noted that the acoustic model is prepared after cepstral mean normalization, so it is positioned near the origin (the point of c0=0 and c1=0). The speech features from distant data will be also normalized around the origin during ASR decoding. It should be emphasized here that a major difference exists in the c0-c1 space based on translation, orientation and shape. Therefore, traditional cepstral mean normalization (CMN) schemes are not expected to improve speech recognition in a meaningful way.

Since the distant data has already been labeled with phoneme-classes, we extract 13 dimensional mean vectors and 13x13 dimensional covariance matrices for each phoneme-class for the distant data. The same information is also obtained for close talk microphone acoustic model. Next, it is possible to represent the distribution and orientation of each phoneme-class elliptically in the c0-c1 plane based on the corresponding mean vectors and covariance matrices as seen in Fig.2.b. Here, different phoneme-classes in the acoustic model are not overlapping or interfering with each other. However, Fig.2.b shows that all classes are strictly overlapping and interfering with each other for the distant data. This is primarily due to the discriminative power of MFCC features is being reduced by the increased distance between the speaker and microphone. As expected, the phonemes of distant data are therefore becoming indistinguishable or unpredictable because of this inter-phoneme-class overlap.

At this point, given knowledge of the specific phoneme classes, we propose to create a mapping strategy for our distant data so that it will be reasonably adapted towards the close talk microphone acoustic model. This adaptation strategy will take each phoneme-class of the input distant test data and project them towards the corresponding locations of the phoneme-classes of the close talk microphone acoustic model. In the Fig.2.c, the bold ellipses correspond to the exact distributions (and alignments) of 'Fricatives' for the acoustic model and input distant test data. In this case, the fricatives for the test data (even after cepstral mean normalization) have an incorrect location, orientation and variance compared to the fricatives of the trained close talk acoustic model. As a result, it will be necessary to perform mean, variance and orientation adaptation.

2.3. Using Principle Component Analysis in PCBA

Next, we create the phoneme-class based adaptation using Principle Component Analysis (PCA). Before applying PCA, we form MFCC dimension pairs (i.e., c0-c1 or c2-c3) in order to select the planes that create the phoneme-class based mapping schemes. After doing this pairing, two dimensional mean vectors and two-by-two covariance matrices for each of our five phoneme-classes are obtained. Then, the eigenvectors of covariance matrices of each phoneme-class of acoustic model and test data are derived. According to PCA, the eigenvectors are the orientations of phoneme-classes and the eigenvalues are the variances of them along the corresponding eigenvector (orientation) [5].

$$G_i(\vec{x}) = \frac{1}{\sqrt{2\pi\,|\boldsymbol{\Sigma}_i|}}\, exp\,(-\frac{1}{2}\,\vec{x}^{T}\boldsymbol{\Sigma}_i^{-1}\vec{x}) \qquad (1)$$

Here, $G_i(\boldsymbol{x})$ represents the Gaussian distribution of the phoneme-class i in the c0-c1 plane, where $\boldsymbol{x}$ is a two dimensional vector [$c0_x$ $c1_x$], and $\boldsymbol{\Sigma_i}$ is the covariance matrix of phoneme-class i. There must be two covariance matrices for each phoneme-class i: (1) for the close talk acoustic model, and (2) for the distant test data. Let u_{model} and v_{model} be the eigenvectors, and λ_{modelU} and λ_{modelV} be the eigenvalues of phoneme-class i of the close talk microphone acoustic model. Next, let u_{data} and v_{data} be the eigenvectors, and λ_{dataU} and λ_{dataV} be the eigenvalues of the distant test data.

Next, we wish to use these eigenvectors to transform phoneme class i distribution of distant data to that of trained acoustic model in terms of orientation and variance in c0-c1 plane (mean adaptation will be discussed later). Note that the directions of the vectors u_{model}, v_{model}, u_{data} and v_{data} have been selected in such a way that they only allow for the least possible angular rotation. Assume again: (1) x_{data} is a two dimensional vector [$c0_x$ $c1_x$] which belongs to a single distant

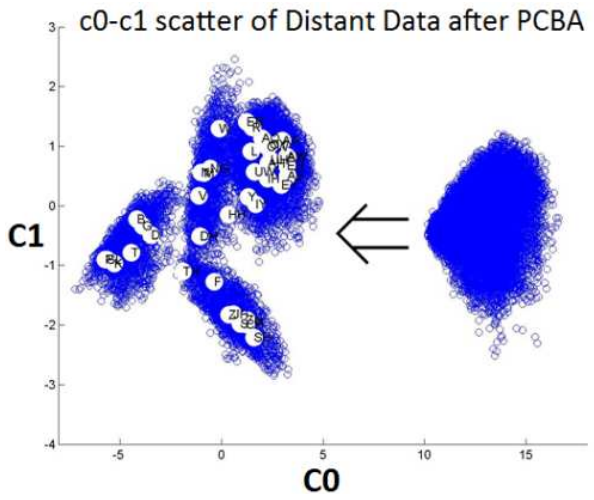

Figure 2: The final appearance of c0-c1 scatter of Distant Data after PCBA at left (together with its appearance before PCBA at right). It can be observed how close to close-microphone acoustic model this new distribution is.

data frame, and (2) that we know this frame belongs to phoneme class i. Now, the new phoneme-class-based adapted vector will be determined as shown in Eq. 2 to 6:

$$coeff_{dataU} = \vec{x}_{data} \cdot \vec{u}_{data} \quad (2)$$

$$coeff_{dataV} = \vec{x}_{data} \cdot \vec{v}_{data} \quad (3)$$

Eq. 2 and 3 are the projections of vector x to the eigenvectors of phoneme class i of the distant data. If we wish to conduct variance adaptation, we must multiply these coefficients according to the eigenvalues as shown below:

$$coeff_{modelU} = coeff_{dataU} \cdot \sqrt{\lambda_{\text{modelU}}/\lambda_{\text{dataU}}} \quad (4)$$

$$coeff_{modelV} = coeff_{dataV} \cdot \sqrt{\lambda_{\text{modelV}}/\lambda_{\text{dataV}}} \quad (5)$$

If variance adaptation is not needed, this step can be skipped. Alternatively, the effect of variance adaptation can be reduced with a desired factor. Finally the phoneme-class-based adapted vector will be:

$$\vec{x}_{Adapted} = coeff_{modelU} \cdot \vec{u}_{model} + coeff_{modelV} \cdot \vec{v}_{model} \quad (6)$$

The main result of phoneme-class based adaptation can be seen Fig. 2, where the c0-c1 cloud (or scatter) located at right belongs to the distant data, and c0-c1 scatter at the left belongs to the corresponding phoneme-class-based adapted version. The phonemes of the close talk microphone acoustic model which are superposed with the proposed phoneme-class-based adapted distribution are also illustrated. Obviously, the distribution of the phoneme-class-based adapted data is very reasonably close to the close talk microphone acoustic model in terms of overall shape, whereas the distribution of raw distant data is not.

It should be noted that, even though these ideas are presented in the c0-c1 plane, all thirteen MFCC dimensions can be paired with each other in order to create new PCBA planes such as c2-c3, c4-c5, etc. After performing this process, a complete phoneme-class based adaptation scheme can be created for every dimension of MFCCs.

2.4. Collaboration of PCBA with MLLR/MAP adaptations

PCBA projects the distant data into a space which is wider and more closely associated with the close-talk acoustic model. A fine phoneme-level adaptation (such as MLLR/MAP adaptation [6]) can follow the proposed phoneme-class level adaption to complete a comprehensive adaptation. Therefore, phoneme-class based adaptation will be viewed as the "primary adaptation", and phoneme-level MLLR or MAP adaptation will be viewed as the "secondary adaption" for distant data. We can summarize this idea in the 'Distant-Speech Adaptation Triangle' shown in Fig. 3. First, MFCCs of the distant data are transformed towards the close microphone acoustic model, and then acoustic model is adjusted towards the phoneme-class-based adapted MFCCs.

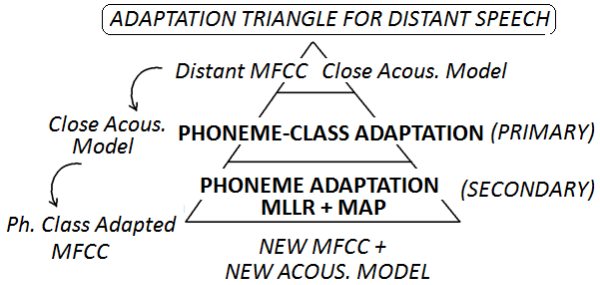

Figure 3: Collaboration of phoneme-class based adaptation (PCBA) and MLLR/MAP adaptation for distant-based data.

3. Experiments

In the experiments, a large 200-seats lecture auditorium was used for the all speeches. The data was obtained from a semester based event of Dep. of Electrical Engineering at the University of Texas at Dallas which is called "Senior Design Day Presentations". In these events, each senior design team is required to present 3-5 min. oral presentation. A distant microphone was located approximately 4-5 meters away from the speaker, and the speaker for each team wore a wireless synchronized close talk microphone. From the Department's archive of recordings, the experiments are focused on a core set of 10 presentations consisting of more than 40 minutes of data. The major challenges for this data are the acoustic mismatch due to distance, room reverberation and large room acoustics, permanent and instantaneous environmental noise. Because the distant microphone is located close to the audience, while it captures the presentations occurring 5 meters away, it is also captures many different types of noises from the audience. The video recording versions of these presentations are also available online: http://www.youtube.com/user/EE1Events1UTD/

Since the proposed phoneme-class based adaptation was developed using context independent (CI) acoustic models, all experiments were performed with CI acoustic models. These acoustic models have only 40 phonemes (monophones), three states per phoneme, and one mixture per state.

3.1. Theoretical Experiments

In this section, the aim is to illustrate the usefulness of phoneme class based adaptation, assuming the perfect phoneme-class information for distant speech data. With this perfect phoneme-class knowledge, this sets an upper bound on

performance improvement. As such, both PCBA and MLLR/MAP based adaptations were performed in supervised processes. The results are expressed in terms of average WERs in Table 2. The rows labeled "DISTANT" are for experiments with raw distant speech test data using close talk microphone acoustic model. "DISTANT+MLLR/MAP" represents the experimental results with MLLR+MAP adaptation applied to distant talk based test data. "PCBA" represents the proposed solution which adapts MFCCs from distant data towards close talk microphone acoustic model. "PCBA+MLLR/MAP" represents PCBA applied to the input test data followed by MLLR/MAP applied to the close-talk trained acoustic models.

Explanation:	WER:
Distant Data	88.9%
MLLR+MAP	78.8%
PCBA	18.1%
PCBA + (MLLR/MAP)	5.6%

Table 2. The results of Theoretical Experiments which are conducted just to show the power of phoneme-class based adaptation (PCBA)

3.2. Practical Experiments

An Artificial Neural Network based phoneme-class decoder was created for use on the distant speech. These ANNs have 60 hidden states, and since only broad classes are needed, only the first four MFCC (c0, c1, c2 and c3) coefficients and also their Δ and $\Delta\Delta$ coefficients are used as input. The ANN produces five channels of information as output. Each of these five channels corresponds to one specific phoneme-class. A frame which belongs to phoneme-class *i* should in general yield a score close to 1 in the *i*th channel and scores close to 0 for all other channels.

Due to their energy characteristics, vowels, fricatives and silence can be predicted up to 75-80% precision and recall for distant speech. However, plosives and the mix-class phonemes (such as nasals, etc.) are partially confused with other phoneme-classes because of their low or mixed energy characteristics, especially for this distant-based data. However, temporal cues can help during their detection, and so they can accurately be predicted up to a 50% precision and recall.

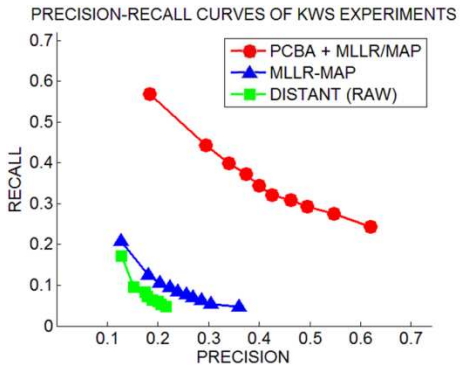

Figure 4: Keyword Spotting Results for Practical experiments with ANN-based phoneme-class detection.

In these experiments, all training and testing data was separated for open evaluations. The performance is illustrated using precision-recall curves of our KWS system which was developed for this distant data scenario. Fig. 4 shows the precision-recall plots for PCBA+MLLR/MAP, MLLR+MAP adaptation alone, and raw distant data.

4. Discussion and Conclusions

When considering the theoretical results conducted under perfect phoneme-class knowledge assumption, it is clear that a significant performance gain achieved between the proposed phoneme-class based adaptation (PCBA) versus MLLR/MAP alone. The basic idea behind PCBA is that for distant based speech, it is not appropriate first to conduct a global MLLR+MAP adaptation. Such a solution will shuffle all the 40 phonemes in the acoustic model throughout the MFCC space. Instead of this scenario, it is more effective first to conduct a "global phoneme-class based adaptation" which locates each phoneme-class to the correct locations reserved within the close talk microphone acoustic model space. After this phoneme-class based adaptation, a "local" MLLR/MAP can be applied to the phoneme-class adapted data which allows for more fine-grain adjustments. Local adaptation is that allows the phonemes only to be arranged within relatively small specific areas which are determined by the phoneme-classes.

When considering results from the practical experiments based on KWS, again it is clear that the proposed PCBA+MLLR/MAP solution achieves superior performance over MLLR/MAP. The reasoning for PCBA superiority can be explained as follows: phoneme-class based adaptation spreads the MFCCs of distant-speech into large areas (just as illustrated in the example c0-c1 plane) in accordance with the close talk acoustic model. This adaptation and any phoneme level adaptation (i.e., MLLR or MAP) which is applied subsequently will be discriminative compared to MLLR/MAP adaptations alone which are directly applied to MFCCs of distant speech. The only advancement offered by MLLR-MAP adaptation alone, which is directly applied to the distant data, is to compress all phonemes in the acoustic model space into a smaller region in accordance with the distant based data features (which in the end reduces the discriminative power of the acoustic model, and therefore limits much performance gains).

Finally, the proposed PCBA method is reasonable because the detection of phoneme-class can be possible at sufficient accuracy for a distant-noisy speech, when individual phoneme detection is not. As the future studies, a new PCBA for context-dependent acoustic models can be created in accordance with context dependent acoustic modeling understandings.

5. References

[1] Demiroglu, C. and Anderson, D. V., "Broad phoneme class recognition in noisy environments using the GEMS", in Proceedings of the ACSSC, 1805-1808 Vol. 2, 2004.

[2] Arslan, L. M. and Hansen, J. H. L., "A minimum cost based phoneme class detector for improved iterative speech enhancement", in IEEE ICASSP-94, II/45-II/48 Vol. 2, 1994.

[3] Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, "Emotion Recognition based on Phoneme Classes", ICSLP-04, 889-892, 2004.

[4] Garofolo, J. S., et al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus" Linguistic Data Consortium, Philadelphia, 1993.

[5] Smith, L. I., "A Tutorial on Principal Components Analysis", Cornell University, USA, Vol. 51 pp. 52, 2002.

[6] Woodland, P.C., "Speaker Adaptation for Continuous Density HMMs: A Review" ISCA Workshop on Adaptation, 11-19, 2001