Frequency Domain Correspondence for Speaker Normalization

Ming Liu, Xi Zhou, Mark Hasegawa-Johnson, Zhengyou Zhang and Thomas S. Huang

IFP, Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL, 61801
[mingliu1,xizhou2,jhasegaw,huang]@ifp.uiuc.edu

Abstract

Due to physiology and linguistic difference between speakers, the spectrum pattern for the same phoneme of two speakers can be quite dissimilar. Without appropriate alignment on the frequency axis, the misalignment will reduce the modeling efficiency resutling in performance degradation. In this paper, a novel data-driven framework is proposed to build the alignment of the frequency axes of two speakers. This alignment between two frequency axes is essentially a frequency domain correspondence of the two speakers. To establish the correspondence, we formulate the task as a global optimal matching problem. The local matching of frequency bins is achieved by comparing the local feature of the spectrogram along the frequency bins. The local feature is actually capturing the local pattern in the spectrogram. Given the local matching score, a dynamic programming is then applied to find the optimal correspondence. Experiments on TIMIT corpus and TIDIGITS corpus clearly show the effectiveness of this method.

1. Introduction

The inter-speaker variation is one of the major challenges to the current automatic speech recognizer. Figure **??** shows some examples of this inter-speaker variation. Due to this variation, the performance of speaker-independent system is generally worse than speaker dependent system. The underline reason of inter-speaker variation is mainly the physiology difference(vocal tract shape and length, etc) and linguistic difference(accent and dialect, etc). Because of these factors, the spectrum pattern for the same phoneme of two speakers can be very different. Without appropriate alignment on the frequency axis, the data variation will dramatically reduce the modeling efficiency and result in performance degradation.

There are many algorithms proposed in the literature to reduce the inter-speaker variation. These methods can be categorized into two classes: model based normalization and feature based normalization. Maximum likelihood linear regression (MLLR)[1] and Maximum A Posterior (MAP)[2], etc are well known model based speaker normalization methods. Vocal tract length normalization (VTLN)[3][4][5][6][7] is a well known algorithm to warp the frequency axis by introducing a warping function. After warping the spectrum, VTLN is able to reduce the inter-speaker variation of different genders and age groups. There are mainly three different warping functions are used in the literature: linear warping, nonlinear warping and piecewise linear warping. In linear warping, one parameter will determine the global warping which may not be able to compensate the total variation of different speakers. Nonlinear warping and piecewise linear warping are proposed to further improve the warping power. In addition to explicitly warping the frequency axis, there is a large amount of research on learning linear transformation for speaker normalization based on maximum likelihood criterion[8]. Surprisingly, it is shown in [7] that the VTLN can be represented as a linear transform in the cepstral domain. All of these normalization methods are essentially maximizing the likelihood of utterance given a model. Instead of maximizing the utterance likelihood, we are trying to find the frequency axis alignment for any two speakers. This alignment is actually a mapping between two frequency axes, in another word a frequency domain correspondence between two speakers. With the right frequency domain correspondence between speakers, the inter-speaker variation can be reduced prior to acoustic modeling procedure which will significantly increase the modeling efficience.

In this paper, we propose a framework of dynamic programming to eatablish the frequency domain correspondence. To construct the metric matrix for dynamic programming, we represent the freqeuncy bin using a descriptor based on the local features which decribes the local pattern in spectrogram. A distance metric is then defined on the pair of descriptors to quantify the similarity between frequency bins. The underline assumption of this method is that the two frequency bin is similar if and only if the local patterns are alike. In this paper, the local feature adopted is the his-

This work was supported by National Science Foundation Grant CCF 04-26627

togram of oriented gradient(HOG). Experimental results on TIDIGITS and TIMIT corpus clearly show the effectiveness of this method.

The paper is organized as follows: Section 2 illustrates the proposed framework. Section **??** shows the experimental results, and conclusions are in Section **??**.

2. Proposed Framework

The correspondence between two frequency axes can be represented by a warping function from one axis to the other.

$$\hat{f} = w(f) \tag{1}$$

where, f is the frequency bin in one axis and the $\hat{f}$ is the corresponded frequency bin in the other. Here, we only consider the discrete frequency axes due to the nature of FFT spectrum. In another point of view, the sequence of pairs $(f, w(f))$ is actually one path in a 2D grid. Figure **??** illustrates one path in the 2D grid is a correspondence between two axes. Here we assume each frequency axis has N discrete frequency bins. Therefore, every possible correspondence is essentially a path in the 2D grid. The problem now becomes which path is the optimal one? If we can define the metric structure between frequency bins, the answer to the problem become clear: the path associated with smallest accumulated distance. Appreantly, the solution to this path finding problem perfectly fit into dynamic programming framework. Now the question is how to define the metric structure between frquency bins.

To construct the metric structure, we represent each freqeuncy bin using a descriptor based on the local pattern in spectrogram. The underline intuition is that the two frequency bin is similar if and only if the local patterns are alike. In our framwork, the local pattern is represented using Hoistogram of Oriented Gradient(HOG) which is a well know local feature from computer vision literature. Before extraction of HOG feature, the speech spectrogram is smoothed to remove the harmonic structure in the spectrogram.

2.1. Smoothed Spectrogram

A spectrogram $S(t, f)$ is a 2D representation of the speech signal based on the short time fourier transform(STFT) analysis. The two axes of spectrogram are time and frequency respectively. For visualizing a given spectrogram $S(t, f)$, the magnitude of a given frequency component f at a given time t in the speech signal is indicated by the darkness or color at the corresponding point. Figure **??** shows a example of colorful spectrograms. There are basically two major cues in spectrogram. One is harmonic cue which is due to the fundamental frequency. The other is formant cue which is due to the vocal tract characteristic. The harmonic cue is more related to speaker characteristic while the formant cue

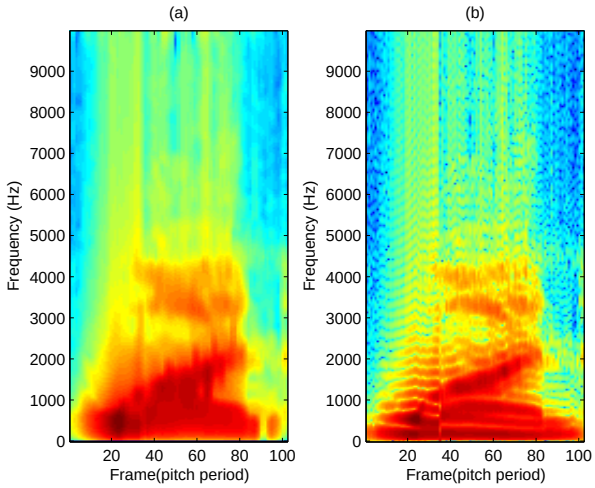

Figure 1: Spectrogram with smoothing vs without smoothing

convey most of the speech content information. In our scenario, the formant cue is the most important information to establish the frequency domain correspondence. To obtain more accurate information from formant cues, we need to smooth out the harmonic structure in spectrogram. In this paper, we adopts a simple algorithm which firstly peak up the spectrum peaks followed with an interpolation to generate the spectral envelope for each frame. To make the spectrogram also smooth along the time line, we use pitch synchronous analysis to generate variable frame length analysis and the pitch is estimated by a open source speech analysis tool – praat[9]. Notice, the estimated pitch period is actually used to define the frame length of STFT. Column (a) and (b) in Figure 1 show the spectrograms with smoothing and without smoothing. As shown in the figure, the smoothed spectrogram preserves the formant location/transition information while smooths out the harmonic structure.

2.2. Histogram of Oriented Gradient

After spectrogram smoothing, the local textual patterns in the spectrogram are captured by a specific local feature – the histogram of oriented gradient(HOG)[10][11] which is a well known feature in computer vision literature. The HOG features are extracted at each a local region centering around every frequency bin f and time t. The HOG basically describes the coarse information about the gradient orientation in a local region of the spectrogram. Figure 2 shows several local patches of a typical spectrogram and their HOGs. Based on some primary experiments, we set the appropriate size of local region to be 10x10 which means 10 frequency bins by 10 frames region centering around position (t, f) in the spectrogram. And the orientation is divided into 8

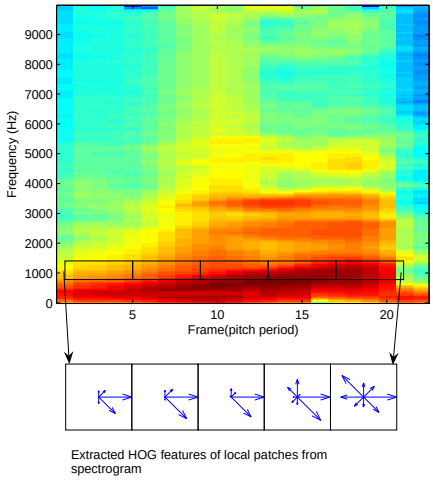

Figure 2: Histogram of Oriented Gradient(HOG) extracted from local patches in spectrogram

equally spaced intervals to cover $[0 - 2\pi)$. The magnitude of gradient at each grid point is added into adjacent intervals according to the distances to the interval boundaries. This will smooth the final histogram which make the HOG feature more robust.

2.3. Simialrity Measure

The extracted HOG features are then normalized into unit length vectors. The set of HOG features along one frequency bin f is denoted as $\{H(:, f)\}$. It is used to describe the local patterns along frequency bin f of all time. Then the simialrity measure between frequency bins is basically the simialrity between two set of HOG features as follows.

$$S(H(:,i), H(:,j)) = \frac{1}{N}\sum_{t=1}^{N} s(H(t,i), H(:,j)) \quad (2)$$

$$s(H(t,i), H(:,j)) = \frac{1}{C}\sum_{k=1}^{C} s(H(t,i), H(t_k,j)) \quad (3)$$

where, $H(:,i)$ and $H(:,j)$ are the HOG feature sets along frequency bin i and j respectively, $S(H(:,i), H(:,j))$ is the similarity measure between these two sets. $s(H(t,i), H(:,j))$ is the similarity between one HOG feature to a set. The similarity between two HOG features $H(t,i)$ and $H(t',j)$ is normalized cross correlation between these two vectors. In our experiments, C is set equal to 3. Notice, the $S(H(:,i), H(:,j))$ is asymmetric between i and j. We can average $S(H(:,i), H(:,j))$ and $S(H(:,j), H(:,i))$ to obtain a symmetric measure. In experiments, we found the performance between asymmetric and symmetric measure is about the same.

2.4. Dynamic Programming Matching

The transition paths is set to be $(i-1, j-1)$, $(i-1, j-2)$ or $(i-2, j-1)$ as illustrated in Figure **??**. The cost for each path is set equal in our implementation. For more details about dynamic matching, Chapter 4.7 in [**?**] provides substantial material on this topic. The boundary conditions of the optimal alignment are listed as follows.

$$w(f_{min}) = f_{min} \quad (4)$$

$$w(f_{max}) = f_{max} \quad (5)$$

where f_min and f_max are starting and ending frequency for our alignment. In this paper, $f_{min} = 0$ and $f_{max} = f_s/2$, and f_s is the sample rate of speech signal.

After obtaining the optimal alignment obtained by dynamic matching, we use following warping to warp one speaker's spectrogram

$$\bar{S}(t,f) = S(t, w(f)) \quad (6)$$

where $S(t,f)$ is the source spectrogram and $w(\cdot)$ is the optimal alignment function, $\bar{S}(t,f)$ is the warped spectrogram.

3. Experiments and Results

A set of experiments are conducted to evaluate the proposed method. First of all, we demonstrate the algorithm can establish the correct correspondence between two speakers. Male speaker FF and girl speaker JM from TIDIGITS corpus are chosen for this demonstration. Based on the sentence 1A.wav(.3sec) of these two speakers, a correspondence is learned by proposed method. Figure 3 show the results. It clearly show the correspondence is able to warp the spectrogram of speaker JM to better match with the spectrogram of speaker FF. The first and second formant have been warped to the correct target position.

Clearly, the proposed method is very effective to reduce the inter speaker variation. To further confirm this conclusion, we also warp the training data of each category to make the training data more compact and result in better modeling efficiency. Table 2 shows the result of warping both training and testing utterances. Compared to the test warping experiment, the WER is further reduced at all conditions. The WER of girl's training and man's testing drops from 21.81% to 1.92% which indicate the great efficience of the proposed method.

4. Conclusion and Future Work

This paper propose a new framework to learn the frequency domain correspondence between speakers which basically find the optimal spectral alignment for the two speakers. Experimental results indicate the effectiveness of this new framework. The success of the HOG feature actually provide a alternative way to perform speech recognition: local feature based method. In the near future, we plant to

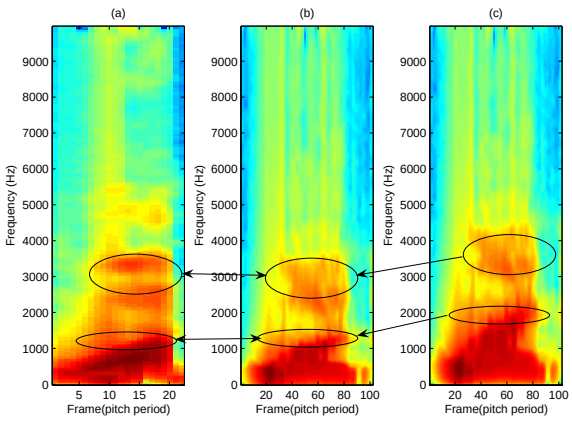

Figure 3: Correspondence between two speakers. Column (a) is the spectrogram of digit one(1A.wav) from the speaker FF in TIDIGITS corpus. Column (c) is the spectrogram of digit one(1A.wav) from the speaker JM in TIDIGTIS corpus. Column (b) is the warped spectrogram of column (c). The two ellipses are used to illustrate the corresponded structures are correctly wrapped by the frequency domain correspondence.

run recognition experiments on large database to verify this method. Also, speech recognition directly after HOG feature extraction will be another promising direction.

5. References

[1] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of the parameters of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–195, 1995.

[2] C. Chesta, O. Siohan, and C. H. Lee, "Maximum a posteriori linear regression for hidden markov model adaptation," in *Proceedings of EUROSPEECH*, 1999, pp. 211–214.

[3] E. Eide et. al., "A parametric approach to vocal tract length normalization," in *Proceedings of ICASSP*, 1996, pp. 346–349.

[4] S. Wegmann et. al., "Speaker normalization on conversational telephone speech," in *Proceedings of ICASSP*, 1996, pp. 339–341.

[5] L. Lee and R. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proceedings of ICASSP*, 1996, pp. 353–356.

[6] M. Pitz and H. Ney, "Vocal tract normalization equals line transformation in cepstral space," *IEEE Trans. on Speech and Audio Processing*, 2003.

WER(%)	man	woman	boy	girl
man	.41	7.59	15.34	22.53
woman	7.54	.22	1.73	2.57
boy	9.97	.75	.47	.77
girl	21.81	2.19	.85	.43

Table 1: Confusion Matrix on TIDIGITS of MFCC Baseline

WER(%)	man	woman	boy	girl
man	.41	6.9	14.19	20.55
woman	7.05	.22	1.63	2.37
boy	9.27	.76	.49	.75
girl	20.23	2.06	.83	.42

Table 2: Confusion Matrix on TIDIGITS of VTLN

WER(%)	man	woman	boy	girl
man	.95	3.76	7.92	13.50
woman	1.72	1.34	1.34	1.60
boy	4.92	.93	.83	.64
girl	1.92	1.20	.65	.44

Table 3: Confusion Matrix on TIDIGITS of Warped MFCC(warp train and test utterances)

[7] L. F. Uebel and P. C. Woodland, "An investigation in vocal tract length normalization," in *Proc. ISCA Europ Conf. on Speech Communication and Technology*, 1999, vol. 6, pp. 2527–2530.

[8] R. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Proceedings of ICASSP*, 1998, pp. 661–664.

[9] Paul Boersma and David Weenink, "Praat: doing phonetics by computer (version 4.3.14)," http://www.praat.org/, 2006.

[10] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[11] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *International Conference on Computer Vision & Pattern Recognition*, Cordelia Schmid, Stefano Soatto, and Carlo Tomasi, Eds., INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, June 2005, vol. 2, pp. 886–893.