# Speaker Normalisation using Centre of Gravity

*R.Sandhya Rani, D. R. Sanand and S. Umesh*
Department of Electrical Engineering,
Indian Institute of Technology, Kanpur, India - 208 016
sandhya.ramisetti@gmail.com,[drsanand,sumesh]@iitk.ac.in

***Abstract*—In this paper, we present a shift based speaker normalisation procedure, where the shift is estimated using spectral centre of gravity(CG) method in each frame instead of conventional ML based methods. The idea is based on the observation that the CG values of two shifted signals also differ by the same shift. The main advantage of this method is that the shift can be estimated in a single step during feature extraction making it computationally efficient as compared to the ML methods. We compare the performance of the proposed method with the conventional ML based VTLN on a phoneme recognition task.**

## I. Introduction

Inter speaker variability is a major source of performance degradation in Automatic Speech Recognition(ASR). This variability can be due to various factors such as speakers: age, gender, speaking style, accent and emotion. Of all these, speakers age and gender add to most of the speaker variability, which in turn are related to the vocal-tract length of the speakers. Accounting for this variability can improve the performance of the recognition system drastically and is known in literature as vocal tract length normalisation (VTLN)[2]. In VTLN, speaker normalisation is brought about by scaling the spectra of speakers enunciating the same sound and most often the scaling relation is assumed to be linear. It is given as:

$$S_A(f) = S_B(\alpha_{AB}\ f) \tag{1}$$

where $\alpha_{AB}$ is the linear scaling or warping factor relating the spectral envelope of the speaker $A$ and $B$ enunciating similar utterances. The above model is also called *linear-scaling model* [1]. In practical ASR systems, as we have access only to the acoustic data and do not have any idea of the speaker's vocal tract length, we do a maximum likelihood based grid search to find the optimal scaling factor $\alpha$ and is given as:

$$\hat{\alpha}_i = \arg\max_{\alpha} Pr(X_i^{\alpha}|\lambda, W_i) \tag{2}$$

where $\alpha_i$ represents scaling factor of the $i^{th}$ utterance $X_i$, given the HMM model $\lambda$ and the transcription $W_i$ for that utterance. This a computationally expensive procedure as the features have to be computed for all values of $\alpha$ before finding the optimal scaling factor for a particular speech utterance. If we can eliminate the need for a grid-based approach, we will gain significantly in the computation during the estimation of optimal scale factor $\alpha$.

For the model in Eq.1, Umesh.et.al[3] suggested that on universally log-warping the spectra of speakers, the scale factor would appear as a translation factor in log-warped domain as:

$$\begin{aligned} s_a(\lambda) &= S_A(f = e^{\lambda}) = S_B(\alpha_{AB}\ e^{\lambda}) \\ &= S_B(e^{\lambda + \ln \alpha_{AB}}) = s_b(\lambda + \ln \alpha_{AB}) \end{aligned} \tag{3}$$

where $s_a(\lambda)$ and $s_b(\lambda)$ represent the spectra of speaker $A$ and $B$ respectively in the log-warped domain. Here the normalisation is brought about by estimating a frequency shift factor instead of a frequency scale factor. If we can estimate the shift in a single step rather than following a maximum likelihood based grid search, we can have a significant gain in computation. With this motivation, in this paper we present a novel shift based speaker normalisation[4] method ,where the shift is estimated using spectral centre of gravity rather than using maximum likelihood based grid search.

The paper is organised as follows: First we give a brief introduction to centre of gravity and the discuss how it could be used for speaker normalisation, followed by results and discussion.

## II. Centre of Gravity

### A. Definition

The classical definition of CG ($\eta$) of a continuous function $f(t)$, is defined as:

$$\eta = \frac{\int_{-\infty}^{\infty} t * f(t)\ dt}{\int_{-\infty}^{\infty} f(t)\ dt} \tag{4}$$

Alternatively CG can be calculated in frequency domain as shown by Stylianou [5].

$$\eta = -\phi^{'}(0) \tag{5}$$

$$= -\phi(1) \tag{6}$$

where $\phi(\omega)$ is the phase spectrum of $f(t)$. This means that the CG of a real signal $f(t)$ is only a function of the first derivative of the phase spectrum at the origin.

### B. Finding delay using CG

Consider two signals, $f_1(t) = \delta(t)$ and $f_2(t) = \delta(t - t_o)$ which differ only by a delay $t_o$. Their corresponding Fourier transforms are given as:

$$\begin{aligned} \delta(t) &\overset{F}{\longleftrightarrow} 1 \\ \delta(t - t_o) &\overset{F}{\longleftrightarrow} e^{-j\omega t_o} \end{aligned}$$

CG calculated using Eq.6

$$\begin{aligned} \eta_1 = -\phi_1(1) &= 0 \\ \eta_2 = -\phi_2(1) &= t_o \\ \text{Therefore, } \eta_2 - \eta_1 &= t_o \end{aligned}$$

So, if the two signals differ by a delay of $t_o$, their CGs also differ by $t_o$. Hence CG can be used to estimate the delay between signals differing by a shift. In the next section we propose our novel shift based approach for speaker normalisation, where the shift is estimated from the spectral CG's of speakers enunciating similar utterances.

## III. Proposed Method

Consider the spectra of a particular phone enunciated by different speakers. According to Eq.1, the spectra of the speaker will be related by a linear scaling relation and according to Eq.3, they appear as shifted versions in the log-warped frequency domain. Since we are interested in finding the shift factor, we will talking only about the log-warped spectra. Let the shifted versions of the spectra of the speakers be represented as: $s_{ph}(\lambda)$, $s_{ph}(\lambda-\tau_1)$, $s_{ph}(\lambda+\tau_2)$ and so on. By taking IDFT, the shift factor appears only in the phase cepstra and is given as:

$$\begin{aligned} s_{ph}(\lambda - \tau_i) \overset{IDFT}{\Longleftrightarrow} & D_{ph}(c)e^{j\tau_i c} \\ & = |D_{ph}(c)|e^{j\phi_{ph}(c)}e^{j\tau_i c} \quad \forall i \end{aligned} \tag{7}$$

where $|D_{ph}(c)|$ is the magnitude and $\phi_{ph}(c)$ is the phase of the phone cepstra respectively. $\tau_i$ is the shift factor, which can be either positive or negative. From Eq.6 the CG of speaker spectra for the phase cepstrum is given as:

$$CG^i_{ph} = \phi_{ph}(1) + \tau_i \quad \forall i \tag{8}$$

Since we are interested only in shift $\tau_i$, the phone phase factor $\phi_{ph}$ has to be eliminated. From Eq.7, we can notice that the phone phase factor $\phi_{ph}$ will be same for all the speakers uttering the same phone. If we can find a reference speaker for the entire population, the shift factor can be estimated with respect to the reference speaker. But in practice, as we do not have a reference or a golden speaker, we use average CG calculated over all speakers as the reference for a particular phone. This is given as:

$$CG^{ref}_{ph} = \frac{\Sigma_{i=1}^{N} CG^i_{ph}}{N} \tag{9}$$

$$= \phi_{ph}(1) + \frac{\Sigma_{i=1}^{N}\tau_i}{N} \tag{10}$$

$$= \phi_{ph}(1) + \tau_{avg} \tag{11}$$

where $CG^{ref}_{ph}$ is the reference CG for a particular phone and $N$ the total number of speakers.

The shift for any spectra is now calculated as the difference of reference CG with its respective CG.

$$shift^i_{ph} = CG^{avg}_{ph} - CG^i_{ph} \quad \forall i \tag{12}$$

$$= \tau_{avg} - \tau_i \quad \forall i \tag{13}$$

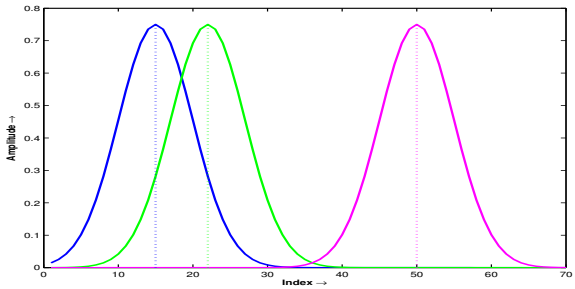


Fig. 1. Original signal

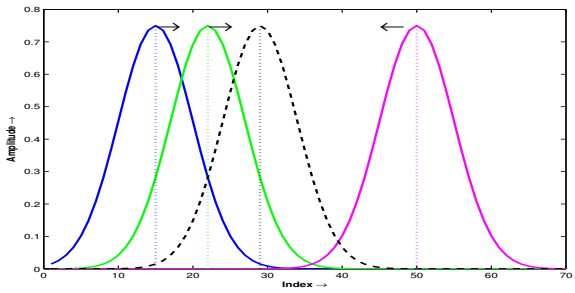


Fig. 2. Normalised signal $shift_i = CG_{avg} - CG_i$

By applying this shift all the spectra move to the reference CG and hence shall be normalised.

The normalisation is illustrated using a synthetic example. Fig.1 depicts three Gaussian signals which are shifted versions of each other. The CG of a Gaussian signal is nothing but its mean(indicated by the vertical dotted line). The reference CG with respect to which the shift will be estimated is the average CG(indicated by the black dotted line). All the spectra shall move towards this reference as shown in Fig.2 and hence are shift normalised.

The normalisation can be equivalently done in the cepstral domain by applying the shift in phase term as:

$$|D_{ph}(c)|e^{j\phi_{ph}(c)}e^{j\tau_i c}e^{j(shift^i_{ph})c} \tag{14}$$

$$|D_{ph}(c)|e^{j\phi_{ph}(c)}e^{j\tau_i c}e^{j(\tau_{avg}-\tau_i)c} \tag{15}$$

Then, the new normalised features are given as: $|D_{ph}(c)|e^{j\phi_{ph}(c)}e^{j\tau_{avg}}$. It can be seen that the speaker specific shift factor $\tau_i$ is nullified.

## IV. Recognition Experiments

We now compare the performance of our proposed shift based speaker normalisation approach with the conventional maximum likelihood based approach for VTLN. The experiments were done on phonemes extracted from the TIMIT database. We have two sets of experiments, one set uses mid-frames (single frame in the centre) of the phoneme and another using all the frames of that particular phoneme. In mid-frame experiments, we used eight vowels in both training and testing, namely *aa, ae, ao, eh, er, ih, iy, ow* and *uw*. In full-frame experiments, we use fifteen vowels both in training and testing, namely *aa, ae, ah, ao, aw, ay, eh, er, ey, ih, iy, ow, oy, uh*

TABLE I
MID FRAME VOWEL RECOGNITION

| Method | Performance |
|---|---|
| Baseline | 57.43 |
| CG-Oracle | 73.64 |

TABLE II
TIMIT FULL FRAME 15-VOWELS

| Method | Performance |
|---|---|
| Baseline | 61.80 |
| VTLN-Oracle | 69.82 |
| CG-Oracle | 66.72 |

and *uw*. For mid-frame experiments, we used 13 dimensional cepstral coefficients containing $c_1, \ldots, c_{12}$ (excluding $c_0$) and normalised log-energy. We used a single emitting state with 8 gaussian mixtures and diagonal covariance. For full-frame experiments, 26 dimensional cepstral coefficients containing $C_1 \ldots C_{12}$ (excluding $C_0$) along with normalised log-energy and the differential coefficients. In either of the cases we performed cepstral mean subtraction.

## V. PERFORMANCE OF PROPOSED METHOD

Table.I shows the performance of the proposed approach on mid-frame data. Here we can not present VTLN performance results as there is only one speech frame per utterance. In this case, we performed an oracle (assume that true transcription is known during testing) experiment on this data to understand the normalisation due to CG. Assuming the true transcription to be known during testing, we are providing information about the reference CG to calculate the necessary shift for normalisation. Once the normalisation is done, we test the normalised utterance with all models and choose the one with maximum log-likelihood as its correct transcription. The results indicate that the method has potential and if performed in the right frame work should yield good results. This experiment also indicates that this is the best possible recognition performance we can obtain.

Another approach to understand the normalisation performance is to measure the separability of the normalised models. One good measure of the separability is F-ratio between models considered pair-wise. It is given mathematically as:

$$fratio_{ab} = \frac{(\mu_a - \mu_b)^2}{(\sigma_a + \sigma_b)/2} \tag{16}$$

where $\mu$ is the mean and $\sigma$ the variance. Higher the F-ratios, better is the separation between models. Tables III and IV show the F-ratios of the baseline and normalised models for mid-frame data. We observe the F-ratios are better for normalised models indicating that the shift estimated using CG of the spectra are indeed doing the right job.

We also performed full-frame vowel recognition experiments in order to understand how the proposed approach works as the confusion in the data increases and also taking into account the co-articulation effects. Table.II shows the performance of full-frame data. Here also, we perform oracle experiments on both VTLN as well as CG methods. We observe that VTLN performs better than CG as expected. This is because, VTLN uses a ML based approach and finds the best warping factor, where as CG method estimates the shift factor required for normalisation in one single step completely eliminating the need for ML based search which was our motivation. Though we are slightly inferior we gain a lot in computation.

## VI. CONCLUSIONS AND FUTURE WORK

In the paper, we proposed a novel shift based speaker normalisation approach using spectral centre of gravity. The main motivation was to eliminate the need for ML based search followed in VTLN to gain computational advantage. Such a method will be of great use for online applications. Though we have only presented preliminary results, which look promising, it requires much more testing before we can come to an assertive conclusion. There are lot of implementation issues that need to be looked at and we are working on them. One such issue is, how to apply the shift? We can apply the shift in the cepstral phase or directly shift the log-smoothed spectrum. Shift in the log-smoothed spectrum can be done by appending zeros in the beginning or at the end, repeating the last or first samples or using a non-uniform scaling method based on the direction of the shift[6]. We also observed that CG is sensitive to DC shifts, which provoked us to investigate in the direction of thresholding the spectra. We are still working on these issues to understand the robustness of the proposed approach.

## VII. ACKNOWLEDGEMENTS

A part of this work was supported by SERC project funding `SR/S3/EECE/0008/2006` from the Department of Science & Technology, Ministry of Science & Technology, India.

## REFERENCES

[1] H. Wakita., "Normalisation of Vowels by Vocal Tract Length and its Application to Vowel Identification. ", *IEEE Trans. ASSP*, Vol.25, No.2, Apr. 1977, pp.183-192.

[2] Li Lee, R. C. Rose, " A Frequency Warping Approach to Speaker Normalisation", *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 1, Jan. 1998.

[3] S. Umesh, L.Cohen, N. Marinovic and D. Nelson. "Scale Transform in Speech Analysis", *IEEE Transactions on Speech and Audio Processing*, Vol.7, No.1, Jan. 1999.

[4] Rohit Sinha and S. Umesh. "Non-Uniform Scaling Based Speaker Normalisation". *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing*, Vol.1, pp 589-592, 2002.

[5] Stylianou.Y, "Removing linear phase mismatches in concatenative speech synthesis", *IEEE Trans. Speech Audio Processing*, Vol.9, pp 232-239, March 2001.

[6] Rohit Sinha, "Front End Signal Processing For Speaker Normalisation", *Ph.D. dissertation, Indian Institute of Technology, Kanpur, India*, June 2004.

TABLE III
FRATIOS BASELINE

| | aa | ae | ah | eh | er | ih | ow | uh |
|---|---|---|---|---|---|---|---|---|
| aa | 0 | 8.5089 | 2.5851 | 9.8083 | 10.3862 | 19.0837 | 3.7922 | 8.4871 |
| ae | 8.5089 | 0 | 4.4791 | 1.1748 | 12.9936 | 5.6635 | 10.2017 | 6.5655 |
| ah | 2.5851 | 4.4791 | 0 | 3.8172 | 8.2609 | 9.6302 | 2.6746 | 3.0485 |
| eh | 9.8083 | 1.1748 | 3.8172 | 0 | 8.9676 | 2.6128 | 8.3729 | 3.2901 |
| er | 10.3862 | 12.9936 | 8.2609 | 8.9676 | 0 | 14.3265 | 11.6121 | 7.9099 |
| ih | 19.0837 | 5.6635 | 9.6302 | 2.6128 | 14.3265 | 0 | 12.7858 | 3.5239 |
| ow | 3.7922 | 10.2017 | 2.6746 | 8.3729 | 11.6121 | 12.7858 | 0 | 3.3048 |
| uh | 8.4871 | 6.5655 | 3.0485 | 3.2901 | 7.9099 | 3.5239 | 3.3048 | 0 |

TABLE IV
FRATIOS CG MODEL

| | aa | ae | ah | eh | er | ih | ow | uh |
|---|---|---|---|---|---|---|---|---|
| aa | 0 | 11.22 | 3.01 | 11.28 | 12.67 | 21.09 | 3.823 | 8.14 |
| ae | 11.22 | 0 | 5.88 | 1.32 | 19.826 | 5.39 | 13.83 | 7.80 |
| ah | 3.01 | 5.88 | 0 | 4.47 | 11.61 | 10.60 | 3.405 | 2.63 |
| eh | 11.27 | 1.32 | 4.47 | 0 | 13.90 | 2.61 | 10.640 | 3.87 |
| er | 12.67 | 19.83 | 11.61 | 13.80 | 0 | 20.58 | 13.798 | 10.31 |
| ih | 21.09 | 5.39 | 10.60 | 2.61 | 20.58 | 0 | 15.811 | 4.87 |
| ow | 3.82 | 13.83 | 3.40 | 10.64 | 13.800 | 15.81 | 0 | 3.75 |
| uh | 8.14 | 7.80 | 2.63 | 3.87 | 10.31 | 4.87 | 3.75 | 0 |