VTLN based on the linear interpolation of contiguous Mel filter-bank energies

Néstor Becerra Yoma[1], Claudio Garretón[1], Fernando Huenupán[2], Ignacio Catalán[1] and Jorge Wuth[1]

[1] Speech Processing and Transmission Laboratory, Department of Electrical Engineering, Universidad de Chile, Santiago, Chile
[2] Department of Electrical Engineering, Universidad de La Frontera, Temuco, Chile
nbecerra@ing.uchile.cl, fhuenu@ufro.cl, jwuth@ing.uchile.cl

Abstract

This paper describes a novel feature-space VTLN method that models frequency warping as a linear interpolation of contiguous Mel filter-bank energies. The presented technique aims to reduce the distortion in the Mel filter-bank energy estimation due to the harmonic composition of voiced speech intervals and DFT sampling when the central frequency of band-pass filters is shifted. The presented interpolated filter-bank energy-based VTLN leads to relative reductions in WER as high as 11.2% and 7.6% when compared with the baseline system and standard VTLN, respectively, in a medium-vocabulary continuous speech recognition task. Also, this new scheme provides significant reductions in WER equal to 7% when compared with state-of-the-art VTLN methods based on linear transforms in the cepstral space. The warping factor estimated here shows more dependence on the speaker and more independence of the acoustic-phonetic content than the warping factor in state-of-the-art VTLN techniques.

Index Terms: speech recognition, speaker normalization, vocal tract length normalization, frequency warping, filter energy interpolation

1. Introduction

Vocal tract length normalization (VTLN) is one of the most popular techniques applied in speech recognition in recent years [1,2-5]. VTLN attempts to reduce the mismatch between training and testing condition in ASR caused by inter-speaker variability as a result of length differences in the human vocal tract. The main idea of VTLN is to align formants between the test speaker and a reference speaker independent or dependent model. VTLN is usually implemented in the front-end by scaling the frequency axis [6-8] or by shifting band-pass filter centre frequencies within filter-banks [2]. Both alternatives can be performed using an optimal warping parameter or factor which is obtained by optimizing a Maximum Likelihood (ML) criterion over the adaptation data via a grid search. As a result, each generated frequency axis or bank-filter per warping factor has to be evaluated [2-5] according to the likelihood of the observed feature vector sequence.

Modeling frequency warping as a linear transform (LT) in the cepstral domain is a strategy that has been followed by some authors [5-7][9-14]. As mentioned in [9], applying VTLN as a LT in the feature-space in cepstral-based ASR presents substantial benefits. For instance, due to the fact that the transform can be applied in the original cepstral features, there is no need to compute the log-filter-bank energies and the discrete cosine transform (DCT) for each evaluated warping factor in the grid search. As a result, the computational load of the VTLN estimation can be dramatically reduced [10].

The LT that models the spectral warping function can be represented in the cepstra [5][11] or in the discrete cepstral space [5-7][9-14]. Those techniques can be interpreted as a particular case of Maximum Likelihood Linear Regression (MLLR) [15-16]. In both groups of techniques the optimal warping factor can be estimated by employing the ML grid search or an analytical gradient-based optimization procedure. For instance, in [6-7][9] the optimization is performed by making use of the ML criterion with an Expectation-Maximization (EM) auxiliary function [17].

The vocal tract frequency response is a continuous function represented by a spectral envelope. However, this frequency response or spectral envelope is evaluated by using two independent discrete sampling processes: first, band-pass filters are modeled with a DFT, which in turn provides a given number of samples within the filter bandwidth; second, the harmonic components in voiced signals sample the vocal tract frequency response at multiples of F0. In Mel filter-banks, which are widely employed in ASR, the filter bandwidths follow the Mel scale. As a consequence, shifting the central frequencies of band-pass filters can introduce perturbations in filter energy estimation due to the discontinuities caused by the DFT and the harmonic structure of voiced signals. This problem is especially acute at low frequencies where the filter bandwidth is narrower according to the Mel scale. For instance, Fig. 1 compares the smoothed spectrum estimated with a moving one-bark bandwidth triangular filter with the reference spectral envelope. As can be seen in Fig. 1.a, the smoothed spectrum obtained with the moving triangular filter is clearly distorted, especially at low frequencies, when compared with the reference spectral envelope. In contrast, the linear interpolation of adjacent Mel filters results in a smoothed spectrum that is much more similar to the spectral envelope (Fig. 1.b). The pitch values within a sentence are highly correlated and we note that the F0 contour does not exhibit large discontinuities [18]. Also, perturbations within a frame result in a likelihood error, which in turn is cumulative on a frame-by-frame basis by definition. Hence, the perturbations due to the harmonic nature of speech will not asymptote to zero as the number of frames increases. Surprisingly, the spectral envelope estimation distortion in VTLN due to the discontinuities caused by the DFT sampling and the harmonic structure of the speech has not been exhaustively addressed in the literature.

In this paper, the warped filter-bank energies are estimated by making use of linear interpolation between contiguous filter energies in the original filter-bank. As a result, the effect of the DFT and harmonic structure of voiced speech intervals is reduced, and hence the perturbation in the spectral envelope estimation is minimized.

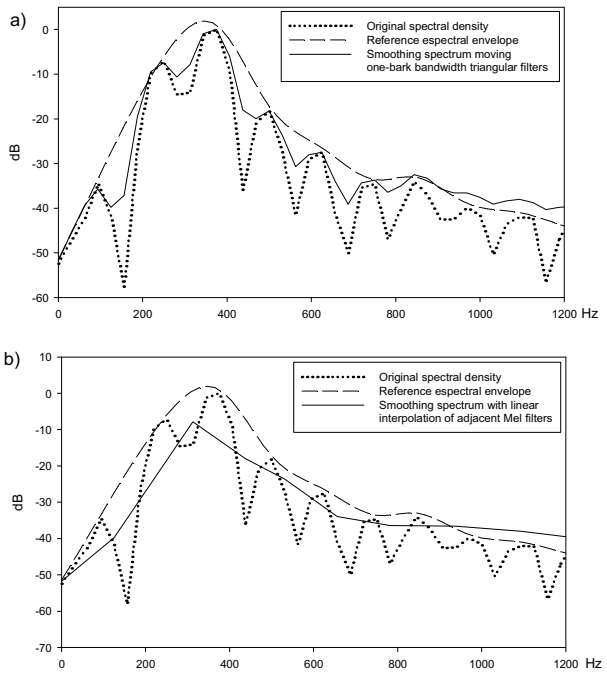

Figure 1: *Spectral density representation of a voiced frame. The original spectral density curve with the harmonic components was estimated by using 256 DFT samples.*

The solution presented here could also be seen as a spectral smoothing method. In this sense, the problem of spectral smoothing has also been addressed by other authors in speech processing. In [19] a method to reconstruct a smoothed time-frequency representation of speech was proposed to reduce the interference caused by the periodicity. In [20] the effect of conventional triangular Mel and uniform-bandwidth filters was investigated in the context of recognition performance for children's speech. Accordingly, it is shown that "differences in spectral smoothing lead to loss in recognition performance with conventional VTLN". However, despite the fact that smoothed spectral estimation is a well known problem in the field of speech science and technology, the VTLN method proposed here has not been found in the literature. Observe that the distortion caused by the harmonic nature of voiced speech in the estimation of warped filter energy is much more evident in the speech of children. Nevertheless, there is no reason to assume that this distortion does not exist in the speech of adults as well. In fact, Fig.1 clearly shows how the harmonic composition of the voice introduces perturbations in the estimation of the spectral envelope.

The contribution of the paper concerns: a) a VTLN model in the filter-bank energy domain based on the interpolation of filter-bank energies (IFE-VTLN); and, b) a comparative analysis of the proposed VTLN method regarding the speaker dependency of the estimated warping factor. It is worth mentioning that the proposed method is also applicable to the interpolation of adjacent filter-bank log energies using a similar mathematical analysis.

As shown later, the interpolated filter-bank energy based VTLN proposed here leads to a linear transform in the cepstral feature-space by approximating the logarithmic function with a first order Taylor series. Experiments with the LATINO-40 database suggest that the presented method can lead to relative reductions in WER as high as 11.2% and 7.6% when compared with the baseline system and standard VTLN, respectively. When compared with state-of-the-art VTLN methods, the proposed ML grid search scheme leads to significant relative reductions in WER equal to 7.0% on average. Finally, the warping factor computed with the VTLN approach described here shows more dependence on the speaker and more independence of the acoustic-phonetic content than the warping factor resulting from standard VTLN and state-of-the-art VTLN methods. This result is observed as a lower gender classification error rate, when the warping factor is employed as a single classifier feature, and as a lower averaged standard deviation per speaker of the warping parameter.

2. Frequency warping and filter energy interpolation

Consider that ω_m is the central frequency of filter m in a filter-bank composed of M filters. Then, $\hat{\omega}_m$ is the warped central frequency of filter m. By using the linear piece-wise warping function proposed in [2], $\hat{\omega}_m$ can be written as:

$$\hat{\omega}_m(\alpha) = \begin{cases} \alpha \cdot \omega_m & \omega_m \le \omega_0 \\ \alpha \cdot \omega_0 + \dfrac{\omega_{\max} - \alpha \cdot \omega_o}{\omega_{\max} - \omega_0}(\omega_m - \omega_0) & \omega_m \ge \omega_0 \end{cases} \quad (1)$$

The energy of filter m at frame i is denoted by $X_{i,m}$. The VTLN method proposed in this paper estimates the energy of warped filter m, $\hat{X}_{i,m}$, as a linear combination of contiguous filter energies in the original filter-bank: if warped filter m is shifted to the left (i.e. $\alpha \le 1$), the warped filter energy is estimated with a linear interpolation between $X_{i,m-1}$ and $X_{i,m}$; and, if warped filter m is shifted to the right (i.e. $\alpha \ge 1$), the warped filter energy is approximated with a linear interpolation between $X_{i,m}$ and $X_{i,m+1}$. Accordingly, $\hat{X}_{i,m}$ is expressed as:

$$\hat{X}_{i,m}(\alpha) = \frac{X_{i,m} - X_{i,q}}{\omega_m - \omega_q}\left[\hat{\omega}_m(\alpha) - \omega_m^{ref}\right] + X_{i,m}^{ref} \quad (3)$$

where,

$$q = \begin{cases} m-1 & \alpha \le 1 \\ m+1 & \alpha > 1 \end{cases} \quad (4)$$

and, $X_{i,m}^{ref}$ and ω_m^{ref} are defined as follows:

$$X_{i,m}^{ref} = \frac{X_{i,m} + X_{i,q}}{2} \quad (5)$$

$$\omega_m^{ref} = \frac{\omega_m + \omega_q}{2} \quad (6)$$

Conventional VTLN is usually implemented by generating a filter-bank for each and every warping factor α to be evaluated. Then, the optimum α is the one that maximizes the likelihood. According to the model presented here, the filter-bank energies for each α to be evaluated can be computed with (3) without the need to run a filter-bank analysis for each α.

By applying the natural logarithm function to (3) and defining $\hat{L}_{i,m}(\alpha) = \log\left[\hat{X}_{i,m}(\alpha)\right]$, filter m log-energy can be written as:

$$\hat{L}_{i,m}(\alpha) = \log\left(\frac{X_{i,m} - X_{i,q}}{\omega_m - \omega_q}\left[\hat{\omega}_m(\alpha) - \omega_m^{ref}\right] + X_{i,m}^{ref}\right) \quad (7)$$

where $\hat{L}_{i,m}(\alpha)$ denotes the warped filter m log energy in frame i. Consider that the observed un-warped MFCC feature vector sequence is denoted with $C = \{C_i\}_{i=0}^{I-1}$, where: $C_i = \{C_{i,n}\}_{n=0}^{N-1}$ corresponds to the frame at instant i, and I is the number of frames; and $C_{i,n}$ denotes the n^{th} cepstral coefficient at frame i, and N is the number of static cepstral parameters. Then, by applying the DCT, $C_{i,n} = \sum_{m=o}^{M-1} L_{i,m} \cos\left(\frac{\pi \cdot n}{M}(m-0.5)\right)$. Consequently, by making use of (7), the n^{th} warped cepstral coefficient at frame i, $\hat{C}_{i,n}$, can be written as [21]:

$$\hat{C}_{i,n}(\alpha) = \sum_{m=o}^{M-1} \hat{L}_{i,m}(\alpha) \cdot \cos\left(\frac{\pi \cdot n}{M}(m-0.5)\right) \quad (8)$$

3. Experiments

Speaker-independent continuous speech recognition results presented in this paper were obtained by using a medium vocabulary task recorded in a clean environment, the LATINO-40 database [22]. This database is composed of continuous speech from 40 Latin American native speakers, with each speaker reading 125 sentences from newspapers in Spanish. The vocabulary is composed of almost 6000 words. In this paper, experiments were conducted using all 40 speakers as test speakers by employing a non-overlapped "leave-four-out" scheme. As a result, ten sub-experiments were carried out with four test speakers each. One HMM was trained per sub-experiment by employing the utterances from the 36 remaining speakers. Consequently, the training data for each sub-experiment corresponds to 4500 utterances. Also, each sub-experiment contains 500 testing utterances, and hence the whole testing database is composed of 10 sub-experiment x 500 utterances per sub-experiment = 5000 utterances.

The band from 300 to 3400 Hz was covered by 14 Mel DFT filters, and at the output of each channel the logarithm of the energy was computed. Thirty-three MFCC parameters (static, delta, and delta-delta coefficients) per frame were computed. Cepstral Mean Normalization (CMN) was also employed. The recognized sentence corresponded to the first hypothesis (the most likely one) within the N-best list obtained from Viterbi decoding. Each triphone was modeled with a

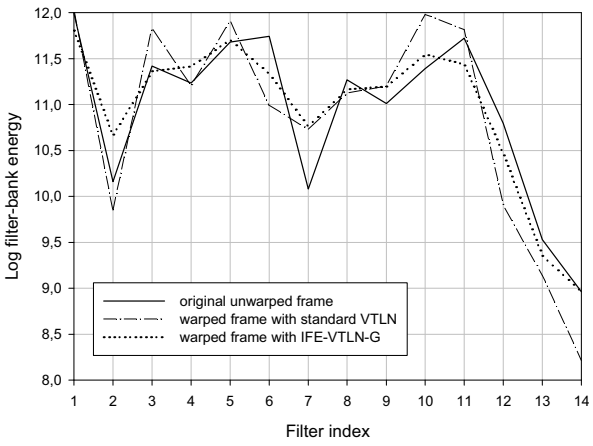

Figure 2*: Spectral envelope of a voiced speech. The utterance corresponds to a male speaker with α = 1.07.*

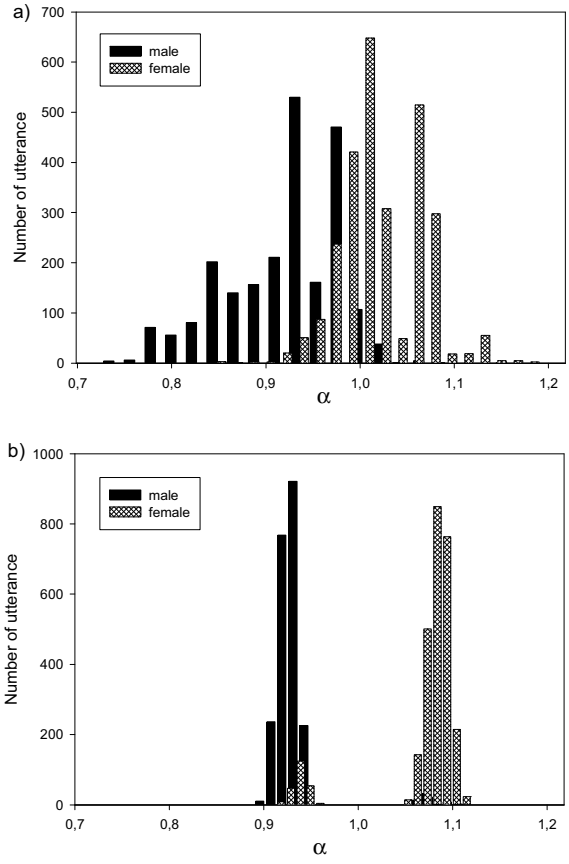

Figure 3: *Histograms of warping factors separated by gender: (a) standard VTLN and (b) the proposed IFE-VTLN-G method.*

three-state left-to-right topology without skip-state transition, with a mixture of eight multivariate Gaussian densities per state with diagonal covariance matrices. The HMMs were trained by using HTK [23] and a trigram language model was employed during recognition. The experiments were conducted by using the recognition engine implemented at the Speech Processing and Transmission Lab., Universidad de Chile. The triphone-based Viterbi algorithm was written by employing ordinary search and pruning techniques in combination with the token passing scheme [24]. The VTLN techniques were applied by estimating the warping factor on an utterance-by-utterance basis with the alignment provided by the best hypothesis in the first Viterbi decoding pass. The baseline system gave a WER equal to 6.42%. The proposed interpolated filter energy model is applied by means of the ML grid search, IFE-VTLN-G. Also, the VTLN technique presented here is compared with the schemes described in [9-10], which are denoted by VTLN-LT1 and VTLN-LT2, respectively. Those methods have been recently proposed in the last few years and successfully model VTLN as a LT in the MFCC domain.

4. Discussion and Conclusions

Figure 2 shows the log-filter-bank energies of the original and warped filter-banks by employing standard VTLN [2] and IFE-VTLN-G as in (3). According to Figure 2, the spectral peaks in the filter-bank energy domain provided by IFE-VTLN-G are similar to those with standard VTLN. However, the differences between the spectral peaks and valleys resulting from IFE-VTLN-G are significantly lower than those provided by standard VTLN. This smoothing effect results from the filter-bank energy interpolation. Also, as proposed here, the filter-bank energy interpolation attenuates the discontinuities caused by the DFT sampling and the harmonic structure of the speech spectrum.

Table 1 shows the WER achieved with the baseline system, standard ML grid search VTLN, IFE-VTLN-G, VTLN-LT1 [9] and VTLN-LT2 [10]. The statistical significance of the differences with respect to IFE-VTLN-G are presented in parentheses. When compared with the baseline system, standard VTLN provides a reduction in WER equal to 3.89%. Also, error rates provided by VTLN-LT1 and VTLN-LT2 are very similar to the one obtained with standard VTLN. This result is consistent with those published in [9-10]. The proposed IFE-VTLN-G scheme leads to relative reductions in WER as high as 11.22%, 7.62%, 6.71% and 7.32%. when compared with the baseline system, standard VTLN, VTLN-LT1 and VTLN-LT2, respectively. This result strongly supports the proposed method.

Table 1. *WER (%) with the baseline system standard VTLN, the proposed IFE-VTLN-G method, VTLN-LT1 and VTLN-LT2. The statistical significances of the differences with respect to IFE-VTLN-G are presented in parentheses.*

	Baseline	Standard VTLN	IFE-VTLN-G	VTLN-LT1	VTLN-LT2
WER-total (%)	6,42 (p<0.003)	6,17 (p<0.036)	5,70	6,11 (p<0.073)	6,15 (p<0.055)

According to Fig. 3 (b), the warping factor estimated with IFE-VTLN-G clearly discriminates between male and female speakers. A similar behavior tends to be observed in Fig. 3 (a). However, the overlap of both populations observed with standard VTLN (and with VTLN-LT1 and VTLN-LT2, although not shown here) is much higher than the one provided by the proposed IFE-VTLN-G scheme. In fact, the gender classification error rates with IFE-VTLN-G, standard VTLN, VTLN-LT1 and VTLN-LT2 are 4.38%, 9.85%, 10.30% and 20.30%, respectively. This result seems to be very interesting when compared with state-of-the-art gender classification technology that can provide accuracies as high as 95%. Vocal tracts in female speakers are usually shorter than in male speakers, which in turns result in higher formant frequencies. Consequently, the lowest gender classification error rate obtained with IFE-VTLN-G suggests that, given a speaker independent HMM, the warping factor estimated with IFE-VTLN-G should depend more on the speaker and be more independent of the acoustic-phonetic content than the warping factor obtained with standard VTLN, VTLN-LT1 and VTLN-LT2. The gender classification error rate was obtained on a sentence-by-sentence basis.

Table 2. *WER (%) by gender with the baseline system, standard VTLN, the proposed IFE-VTLN-G scheme, VTLN-LT1, VTLN-LT2 and VTLN-LT3.*

WER	Baseline	Standard VTLN	IFE-VTLN-G	VTLN-LT1	VTLN-LT2
Male speakers	6,93%	6,65%	6,42%	6,66%	6,30%
Female speakers	6,01%	5,82%	5,11%	5.66%	6,04%

Table 2 presents the WER provided by the baseline system, standard VTLN, IFE-VTLN-G, VTLN-LT1 and VTLN-LT2 separated by gender. When compared with the baseline system, IFE-VTLN-G provides a much higher reduction in WER with female speakers than male speakers (14.98% and 7.36%, respectively). This result strongly supports the hypothesis formulated here and must be due to the fact that female speakers show more separated harmonics in the frequency axis than male speakers. As a consequence, the reduction of the discontinuities due to the harmonic structure of the speech is more relevant for female than male speakers. In contrast, the reductions in WER provided by standard VTLN and VTLN-LT1 with male and female speakers (4.04% vs 3.16% and 3.9% vs 5.8%, respectively) are similar. In the case of VTLN-LT2 significant reductions of WER were observed mainly with male speakers.

5. Acknowledgements

This research was funded by Conicyt-Chile under grants Fondecyt 1100195 and 11110391, and Team Research in Science and Technology ACT 1120.

6. References

[1] A. Andreou, T. Kamm and J. Cohen. "Experiments in Vocal Tract Normalization". Proc. CAIP Workshop: Frontiers In Speech Recognation I. 1994.

[2] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," IEEE Trans. Speech Audio Process., 6(1), pp. 49–60, 1998.

[3] S.Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech". ICASSP 1996 , pp. 339–341, 1996.

[4] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," ICASSP 1996 , pp. 346–349, 1996.

[5] M. Pitz and H. Ney, "Vocal Tract Normalization Equals Linear Transformation in Cepstral Space," IEEE Trans. on Speech and Audio Process., 13(5-2), pp. 930-944, 2005.

[6] S. Wang, X. Cui, and A. Alwan, "Speaker adaptation with limited data using regression-tree-based spectral peak alignment," IEEE Trans. Audio, Speech, Lang. Process., 15(8), pp. 2454-2464, 2007.

[7] Xiaodong Cui and A. Alwan, "Adaptation of children's speech with limited data based on formant-like peak alignment," Computer speech & language, 20(4), pp. 400-419, 2006.

[8] A. Acero and R. Stern. "Robust speech recognition by normalization of the acoustic space". ICASSP 1991, 893-896 vol.2. 1991.

[9] S. Panchapagesan and A. Alwan, "Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC," Computer Speech and Language, 23(1), pp. 42-46, 2009.

[10] S. Umesh, A. Zolnay and H. Ney, "Implementing frequency-warping and VTLN through linear transformation of conventional MFCC." Interspeech 2005, pp. 269-272, 2005.

[11] J. McDonough, T. Schaaf and A. Waibel. "Speaker adaptation with all-pass transforms," Speech Communication Special Issue on Adaptive Methods in Speech Recognition, 2004

[12] D.R. Sanand, R. Schlüter, and H. Ney, "Revisiting VTLN Using Linear Transformation on Conventional MFCC," Interspeech 2010, pp. 538-541, 2010.

[13] T. Claes, I. Dologlou, L. Bosch and D. Compernolle, "A novel feature transformation for vocal tract length normalization in automatic speech recognition," IEEE Trans. on Speech and Audio Process., 11 (6), 603–616, 1998.

[14] D. Giuliani, M. Gerosa and F. Brugnara, "Improved automatic speech recognition through speaker normalization," Computer Speech and Language, 20(1), pp.107-123, 2006.

[15] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, 9, pp. 171–185, 1995.

[16] M.J.F Gales."Maximum Likelihood Linear Transformations for HMM-based Speech Recognition". Computer Speech and Language, 12, pp. 75–98, 1998

[17] T. K. Moon, "The expectation-maximization algorithm", IEEE Signal Processing Magazine, pp. 47–60, 1997.

[18] D. Joho, M. Bennewitz and Behnke, S., "Pitch Estimation using Models of Voiced Speech on Three Levels", ICASSP 2007, vol IV, pp. 1077-1080, 2007.

[19] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," ICASSP 1997, vol.2, pp.1303-1306, 1997.

[20] S. Umesh, R. Sinha, "A Study of Filter Bank Smoothing in MFCC Features for Recognition of Children's Speech," IEEE Trans. on Audio, Speech, and Language Process., vol.15, no.8, pp. 2418-2430, 2007.

[21] N.B. Yoma, C. Garreton, F. Huenupan, I. Catalan, and J. Wuth, "On Reducing Harmonic and Sampling Distortion in Vocal Tract Length Normalization". IEEE Transactions on Audio, Speech and Language Processing, vol. 21, no. 1,pp. 101-121, 2013.

[22] J. Bernstein et al., LATINO-40 Spanish Read News, Linguistic Data Consortium, Philadelphia, 1995.

[23] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P.C. Woodland. The HTK Book, Cambridge University Engineering Department, Cambridge, UK, 2006.

[24] S.J. Young, N.H. Russell, and J.H.S. Thornton. "Token passing: a simple conceptual model for connected speech recognition systems." Technical Report CUED/F-INFENG/TR.38, Cambridge University Engineering Department, 1989.