VTLN ADAPTATION FOR STATISTICAL SPEECH SYNTHESIS

Lakshmi Saheer[1,2], *Philip N. Garner*[1], *John Dines*[1], *Hui Liang*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Federale, Lausanne (EPFL), Switzerland

ABSTRACT

The advent of statistical speech synthesis has enabled the unification of the basic techniques used in speech synthesis and recognition. Adaptation techniques that have been successfully used in recognition systems can now be applied to synthesis systems to improve the quality of the synthesized speech. The application of vocal tract length normalization (VTLN) for synthesis is explored in this paper. VTLN based adaptation requires estimation of a single warping factor, which can be accurately estimated from very little adaptation data and gives additive improvements over CMLLR adaptation. The challenge of estimating accurate warping factors using higher order features is solved by initializing warping factor estimation with the values calculated from lower order features.

Index Terms— Statistical Speech Synthesis, Vocal Tract Length Normalization, Adaptation.

1. INTRODUCTION

Recent advances in the field of statistical speech synthesis [1], have considerably reduced the gap between basic techniques used in automatic speech recognition (ASR) and text to speech (TTS). Feature types, feature dimensionality, duration and pitch modeling are a few of the key differences between the recognition and synthesis models [2]. To augment the ASR models, speech synthesis also uses a duration model by way of the hidden semi-Markov models (HSMM). The general aim of this research is to combine the features used for ASR and for TTS [3]. One particular focus is the use of ASR based adaptation to control the characteristics of a synthesized voice [4]. Vocal tract length normalization (VTLN) is one of the techniques which can be used to remove speaker specific characteristics in order to build improved average voice models. This paper investigates the use of VTLN for adaptation in statistical speech synthesis.

Speaker adaptation is a technique for transforming the model parameters to match the speaker characteristics of a target speaker. Speaker adaptive training helps to build improved speaker independent models by transforming the model parameters and removing speaker characteristics for each speaker in the training data. The most common adaptation techniques are MLLR (Maximum Likelihood Linear Regression), CMLLR (Constrained MLLR), SMAPLR (Structural Maximum A Posteriori Linear Regression) and CSMAPLR (Constrained SMAPLR). Speaker normalization, on the other hand, transforms the feature vectors rather than the model parameters. Feature transformation can be shown to be analogous to model transformation [5]. Usually, speaker adaptation techniques perform affine transformations on the mean and variance of the probability density functions of the HMM states. This can be accomplished to some extent with normalization techniques like VTLN. The main advantage of feature normalization is that the number of parameters to be estimated from the adaptation data is generally smaller compared with the standard model based adaptation techniques. Hence, adaptation can be carried out with very little adaptation data.

VTLN is inspired from the fact that the vocal tract length varies across different speakers. This length varies from around 18 cm in males to around 13 cm in females. The formant frequency positions are inversely proportional to the vocal tract length. This causes variation of around 25% in the formant center frequencies among speakers. Hence, the feature vectors extracted from the speech of different speakers can be normalized to represent an average vocal tract.

Mel-generalized cepstral coefficients (MGCEP) [6] are one of the best known features for statistical speech synthesis. The generalized cepstral analysis method can be viewed as a unified approach to the cepstral and the linear prediction methods, in which the model spectrum varies continuously from all-pole to cepstral according to the value of an analysis parameter, γ. This feature extraction technique involves optimization of two parameters (namely, α and γ). The warping parameter, α, determines the frequency warping of the cepstra. The frequency transformation used in MGCEP extraction is the bilinear transform, which is an all-pass transform. This same all-pass transform is commonly employed in VTLN [7]. Hence, in this work, these two transforms are combined, and VTLN is applied at the feature extraction step. In the context of MGCEP features, VTLN can be considered as finding the optimal warping factor for each speaker.

In this paper, the implementation of VTLN as a bilinear transform for ASR is considered. Its relationship with MGCEP features is reviewed, and solutions to some challenges involving maximum likelihood warping factor estimation for higher dimensional features are presented. An equivalent synthesis system is described that uses a bilinear transform based VTLN. Both objective and subjective evaluations are presented, followed by some discussion and conclusions supporting the use of VTLN.

2. STATISTICAL SPEECH SYNTHESIS

The HMM-based speech synthesis system (HTS) [1] models spectrum, F0 and duration simultaneously in the unified framework of HSMM. In the training stage, the output vector of the HSMM consists of a spectrum part and an F0 part. In the synthesis stage, arbitrary text is converted to a context-dependent label sequence. A sentence HSMM is constructed by concatenating corresponding HSMM models. A state sequence that maximizes the probability for the given sentence is determined. Then a speech parameter vector sequence is generated for this state sequence by speech parameter generation algorithms. Finally, a speech waveform is generated from the speech parameter vector sequence. Adaptation techniques are used in the same way in both TTS and ASR. Speaker adaptive models are built using the adaptation techniques that remove the influence

of speaker characteristic from the training data. During synthesis, models are adapted to a target speaker and thus, synthesizing speech of this speaker using the adaptation data.

It has been shown that the speaker adaptive models can perform better than speaker independent models. Techniques like CMLLR have been used for building speaker adaptive models for TTS. This technique requires many parameters to be estimated in the transform and hence requires more adaptation data during synthesis. Techniques like VTLN have a single parameter to be estimated and hence, requires less adaptation data. CMLLR [8] is a powerful model based adaptation technique that can be shown to be equivalent to a feature transform [5]. VTLN in combination with CMLLR has the potential to perform better, even when there is little adaptation data or when using lower dimensional features for synthesis. These qualities of VTLN can be inherited by TTS, but the application of VTLN to TTS involves additional challenges like estimating warping factors from higher order features and using VTLN with the synthesis features like MGCEP. These challenges are addressed in the following sections.

3. VTLN BASED ADAPTATION

VTLN tries to normalize the position of the formant peaks by warping the spectrum to represent an average vocal tract. The components involved in this technique are:

- A Warping function (linear, piecewise linear, non-linear, bilinear, etc.)
- A Warping factor (α for bilinear transform)
- An Optimization criteria (MAP, ML, MGE, etc.)

One of the main advantages of VTLN is that the warping factor can be reliably estimated even with a single adaptation sentence for each test speaker. We also note an advantage of using bilinear transform based VTLN is that it can be embedded into the frequency warping of the MGCEP features.

3.1. Bilinear Transforms

The bilinear transform of a simple first order all-pass filter with unit gain can be represented as:

$$\psi_\alpha(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} = e^{-j\beta_\alpha(\omega)}, |\alpha| < 1 \quad (1)$$

where α is the warping factor. The warping performed by this function is shown in Figure 1. It can be observed that, for a specific value of $\alpha = 0.42$, this transform can approximate the mel-scale warping.

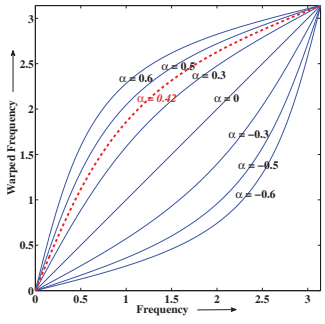


Fig. 1: Bilinear Transform

Bilinear transforms are established as a means of approximating common VTLN transforms [9], and also as a means of performing common frequency warps [6]. In the present study, these advantages are combined with the fact that the bilinear transform can be represented as a linear transform in the cepstral domain.

3.2. VTLN with MGCEP

The feature normalization can be represented as a linear function that transforms the model parameters [5]. A common representation of this linear function is the matrix transformation. The cepstral features are warped using the matrix representation as follows:

$$c_\alpha = S_\alpha c, \quad (2)$$

where α is the warping parameter applied to the unwarped cepstra, c, in order to yield warped cepstra, c_α. S_α is the matrix transformation. It can be shown that the following matrix transformation for MGCEP feature can be derived from the MGCEP recursion [6].

$$S_\alpha = \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{M-1} \\ 0 & 1-\alpha^2 & 2\alpha(1-\alpha^2) & \cdots & M\alpha^{M-1}(1-\alpha^2) \\ 0 & -\alpha(1-\alpha^2) & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & (-1)^{N-1}(1-\alpha^2)\alpha^{N-1} & \cdots & \cdots & \cdots \end{bmatrix}$$

It can also be shown that the elements of this matrix can be estimated using the following recursive formula for $k > 1$ and $l > 1$

$$S_\alpha(k,l) = S_\alpha(k-1,l-1) + \alpha[S_\alpha(k,l-1) - S_\alpha(k-1,l)]$$

3.3. Estimating Warping Parameters

A bilinear transform based VTLN has been implemented in the MGCEP feature extraction with a maximum likelihood (ML) optimization criteria. MGCEP already has a bilinear warping with $\alpha = 0.42$ approximating the mel-scale frequency warping. Another stage of bilinear transform can be cascaded with the existing one to accommodate the VTLN warping. It has been shown [10] that the combination of two bilinear transforms with warping factors α_1 and α_2 is equivalent to a bilinear transform with single warping factor given by:

$$\alpha = \frac{\alpha_1 + \alpha_2}{1 + \alpha_1\alpha_2} \quad (3)$$

3.3.1. Conventional ML based VTLN Estimation

The bilinear transform based warping function has only a single variable α as the warping factor which is representative of the ratio of the vocal tract length of the speaker to the average vocal tract length. The brute force way of computing the warping factor for each speaker is the ML based grid search technique. Maximum likelihood optimization is given by [11]:

$$\hat{\alpha}_{s1} = \arg\max_\alpha \Pr\left(X_{\alpha_{s1}} | M, W_{s1}\right) \quad (4)$$

where $X_{\alpha_{s1}}$ represents the features warped with the warping factor α_{s1}, which is the warping factor for speaker "s1". M represents the model and W_{s1} represents the transcription corresponding to the data from which the features are extracted for speaker "s1". $\hat{\alpha}_{s1}$ represents the best warping factor for the same speaker.

4. EVALUATION OF VTLN FOR SYNTHESIS

The adaptation data is used to estimate the warping factor for each target speaker. This warping factor can be used to adapt the synthesized speech for each speaker. Although VTLN cannot capture the entire characteristics of the speaker with the warping factor, at least

the gender characteristics can be accurately represented. This enables the synthesized voice to sound closer to the voice of the target speaker. Hence, VTLN has the potential to improve adaptation using little adaptation data along with other adaptation techniques like CMLLR.

4.1. Experiments

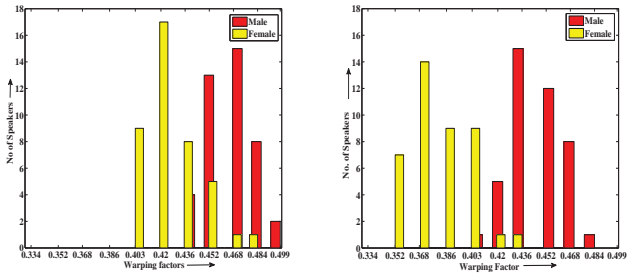
An ML based grid search technique for VTLN is used in this paper. In the training phase, warping factors are initially estimated using grid search and the average voice models are iteratively trained by re-estimating the warping factors until convergence of the model likelihood on the training data. The same grid search technique is used to estimate the best warping factor for each test speaker using the available adaptation data from the corresponding speaker. The grid search for the warping factors is performed with $\alpha_1 = 0.42$, and $-0.1 \leq \alpha_2 \leq 0.1$ with a step size of 0.02. The two transforms are combined using Equation 3.

Full context HSMM models are trained using the HTS 2.1 [12] scripts and are then converted to HMM models. The wall street journal (WSJ0 SI-84) database is used to built the speaker independent models. The HMM toolkit (HTK) is used to align the warped feature vectors with the full context labels and, hence, calculate the log likelihood scores. These scores are compared to obtain the best warping factor for each speaker during training. The statistical models are retrained using features normalized using the estimated warping factor for each speaker in the training data. The warping factor estimation is iterated twice to build better average voice models.

4.2. Issues of dimensionality

HMM based speech synthesis systems require modeling of higher order features when compared to the speech recognition models. It was observed that the warping factor calculation was not successful with the higher (25^{th} or 39^{th}) order features, but worked with lower (12^{th}) order features. Similar observations can be seen in the literature [13, 14]. The work of [14] uses VTLN along with the MCEP (mel-cepstral) features in a similar way but restricting the estimation of the warping factor from only first few cepstral coefficients. The authors experimentally find that using only first 4 coefficients of cepstral features gives better average voice in synthesis. However, the approach taken by [14] is inaccurate due to the fact that the convergence of likelihood values is not guaranteed by warping the entire feature vector with the warping factors estimated from a few cepstral coefficients.

The failure of warping factor estimation for higher order features can be attributed to the presence of excitation harmonics, which could lead to a large likelihood mismatch even for a small warping. It follows that the use of higher order features approaching 25^{th} or 39^{th} order MGCEP should be avoided when estimating warping factors. Instead, the warping factor estimated from the 12^{th} order features can be used as the seed values during the iterative VTLN training for higher dimension features. It is observed that once a good initialization is given, the second iteration of VTLN training is able to estimate good warping factors even for higher order features. This phenomena is illustrated in Figure 2. It can be observed from the figure that the distribution for warping factors estimated from the 25^{th} order has large overlap for male and female speakers with no proper separation of warping factors for female speakers. A more distinct bimodal distribution is observed when the warping factors are initialized with values estimated from the 12^{th} order features.



(a) 25^{th} order system (b) 25-12 system

Fig. 2: Warping factors estimated from 25^{th} order features. The 25-12 system initializes the features with the warping factors estimated from 12^{th} order features. Both graphs have same range for X-axis.

4.3. Evaluation Metrics

Objective evaluation of the synthesized speech is performed using a mel-cepstral distortion (MCD) measure, which is the average Euclidean distance between reference and synthesized mel-cepstral feature vectors. This can be considered to be equivalent to log-spectral distortion according to Parseval's theorem. The convergence of log-likelihood scores during training is presented as a cue for the improvement in the average voice model. A standard adaptation technique (CMLLR) is used to compare the results of VTLN. VTLN together with CMLLR is also synthesized to enable possible additive improvements.

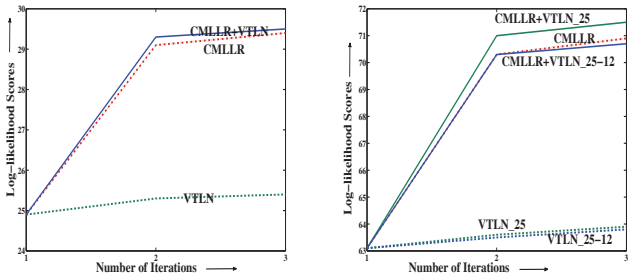
Subjective evaluation of the synthesized speech was conducted to determine mean opinion scores (MOS) for naturalness and speaker similarity. The naturalness was scored on a five point scale ranging from 1 to 5, where 1 represents completely unnatural speech and 5 completely natural speech. Speaker similarity was also rated on a five point scale from 1 to 5, where 1 denotes speech from a totally different speaker and 5 denotes speech from exactly same speaker. Subjective evaluations were conducted on 60 randomly picked sentences from 10 different systems. 19 listeners were presented with the 60 sentences, randomly sorted to avoid any bias due to listening order. The 25^{th} order system for VTLN, CMLLR and CMLLR combined with VTLN were tested with different amounts of adaptation data. These systems were also compared with their respective 25-12 counterparts, where the warping factors were initialized from 12^{th} order and re-estimated using 25^{th} order features.

4.4. Results and Discussion

The experiments are performed on the MGCEP features with the analysis parameter, γ, equal to zero and with two different feature orders, 12 and 25. Evaluations are performed on the incremental speaker adaptive (S4-C3) data set of the WSJ Nov93 test specifications. The results of objective evaluations are plotted as graphs. The log-likelihood scores increase with multiple iterations of each adaptation technique as shown in Figure 3. The MCD results for VTLN based feature adaptation are given in Figure 4. The feature order 25-12 represents the 25^{th} order features initialized with a warping factor estimated from 12^{th} order features. It can be seen that CMLLR leads to additive improvements in performance in combination with VTLN. It can be seen that the average voice model trained with CMLLR and VTLN has better convergence during training and higher MCD during synthesis indicating that it should be a better

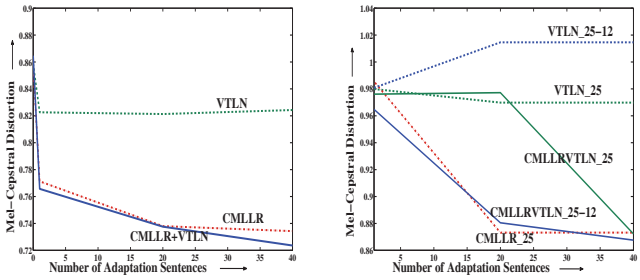
average voice model. It can be observed for 25^{th} order features that VTLN and CMLLR combined with VTLN have lower MCD than CMLLR when only a single adaptation sentence is available. Also, the adapted speech with VTLN in combination with CMLLR gives lower MCD for any amount of adaptation data, suggesting that VTLN can contribute to improvement of the synthesized speech.

Results for subjective evaluations are shown in Figure 5, which shows MOS for naturalness and speaker similarity. Subjective tests were conducted on 10 different systems. These include VTLN, CMLLR and CMLLR+VTLN for 25^{th} order and 25-12 systems with adaptation using 1 and 40 sentences. It is observed that VTLN systems are preferred over other systems for the naturalness cue. Also, VTLN combined with CMLLR is preferred as having better similarity to the voice of the original speaker. The subjective evaluations as such only have limited statistical significance since it is observed that the CMLLR system was not preferred at all for naturalness or speaker similarity. But, these scores support the results from objective evaluations emphasizing the fact that VTLN can lead to additive improvements when combined with CMLLR.



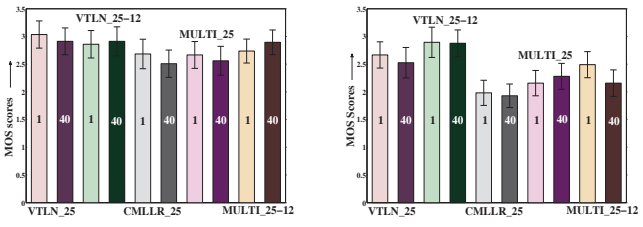
(a) 12^{th} order features (b) 25^{th} order features

Fig. 3: Log-likelihood scores during training.



(a) 12^{th} order features (b) 25^{th} order features

Fig. 4: Mel-Cepstral Distortion for synthesized speech.



(a) MOS for speaker similarity (b) MOS for naturalness

Fig. 5: MOS for naturalness and speaker similarity of synthesized speech. 1 and 40 represents number of adaptation sentences. MULTI represents the combination of VTLN and CMLLR adaptation techniques.

5. CONCLUSIONS

This research has successfully implemented VTLN based adaptation for statistical speech synthesis and incorporated the warping at the feature extraction stage of MGCEP features. It was observed that the VTLN parameters can be accurately estimated from much less adaptation data, as little as a single sentence. VTLN adaptation can estimate the correct gender characteristics of the speech with a single adaptation sentence, and hence the adapted sentence sounds more similar to the original speaker. The warping factor estimation for higher order features can be improved by initializing with values estimated from lower order features. It was also observed that VTLN gives additive improvements when combined with CMLLR adaptation.

6. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Junichi Yamagishi for his assistance with the HTS system training scripts. The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project).

7. REFERENCES

[1] A. W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Proc. of ICASSP*, Hawaii, USA, 2007, pp. 1229–1232.

[2] J. Dines, J. Yamagashi, and S. King, "Measuring the gap between HMM-based ASR and TTS," in *Proc. of Interspeech*, UK, 2009, pp. 1391–1394.

[3] J. Dines, L. Saheer, and H. Liang, "Speech recognition with synthesis models by marginalising over decision tree leaves," in *Proc. of Interspeech*, UK, 2009, pp. 1395–1398.

[4] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 66–83, 2009.

[5] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Transactions on Speech and Audio Processing,*, vol. 13, pp. 930–944, 2005.

[6] K. Tokuda, "Mel-generalized cepstral analysis —a unified approach to speech spectral estimation," in *Proc. of ICSLP*, 1994, pp. 1043–1046.

[7] J. W. McDonough, *Speaker Compensation with All-Pass Transforms*, Ph.D. thesis, John Hopkins University, 2000.

[8] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language,*, vol. 12 (2), pp. 75–98, 1998.

[9] D. Pye and P. C. Woodland, "Experiments in speaker normalisation and adaptation for large vocabulary speech recognition," in *Proc. of ICASSP*, 1997, pp. 1047–1050.

[10] A. Acero, *Acoustical and environmental robustness in automatic speech recognition*, Springer, US, 1993.

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

[12] J. Yamagashi, H. Zen, T. Toda, and K. Tokuda, "Speaker independent HMM based speech synthesis system - HTS-2007 system for blizzard challenge 2007," in *Proc. of Sixth ISCA Workshop on Speech Synthesis*.

[13] T. Emori and K. Shinoda, "Rapid vocal tract length normalization using maximum likelihood estimation," in *Proc. of Eurospeech*, 2001, pp. 1649–1652.

[14] M. Hirohata, T. Masuko, and T. Kobayashi, "A study on average voice model training using vocal tract length normalization," *IEICE Technical Report,*, vol. 103 (27), pp. 69–74, 2003, In Japanese.