A STYLE CAPTURING APPROACH TO F0 TRANSFORMATION IN VOICE CONVERSION

Gopala Krishna Anumanchipalli[†‡], Luís C. Oliveira[‡], Alan W Black[†]

[†]Language Technologies Institute, Carnegie Mellon University, USA
[‡] L^{2}F Spoken Language Systems Lab, INESC-ID / IST Lisboa, Portugal
{gopalakr,awb}@cs.cmu.edu, lco@l2f.inesc-id.pt

ABSTRACT

In this paper, we present a new approach to F0 transformation, that can capture aspects of speaking style. Instead of using the traditional 5ms frames as units in transformation, we propose a method that looks at longer phonological regions such as metrical feet. We automatically detect metrical feet in the source speech, and for each of source speaker's feet, we find its phonological correspondence in target speech. We use a statistical phrase accent model to represent the F0 contour, where a 4-dimensional TILT representation is used for the F0 is parameterized over each feet region for the source and target speakers. This forms the parallel data that is the training data for our transformation. We transform the phrase component using simple z-score mapping. We use a joint density Gaussian mixture model to transform the accent contours. Our transformation method generates F0 contours that are significantly more correlated with the target speech than a baseline, frame-based method.

Index Terms— Prosody Transformation, Metrical Foot, Voice Conversion, F0

1. INTRODUCTION

Voice conversion aims to convert the speech from one speaker and make it sound like another speaker. It has been an active research topic for the last two decades with primary focus on source modifications (energy, pitch etc.,) and filter modifications (for the vocal tract). These approaches took a rather low level perspective on speech ignoring higher level aspects that are otherwise shown to be important. [1] finds that professional impersonators capture aspects of speech style, particularly the rhythm, intonation and stress patterns across words and phrases. However, this aspect of speech style capture is relatively less studied for voice conversion. [2] notes that the biggest challenge at this stage for voice conversion algorithms is the control (modeling, mapping and modification) of the speaking style of a speaker.

In this work, we take a step closer in the direction of capturing the speaking style of a target speaker. Prosody is the general area of study that deals with the speaking style and intonation is perhaps the most important aspect that it is related to. While the correlates of speaking style also exist in duration and phrasing, in this work we focus specifically on the F0 contour. Within the contour, the speaking style is manifested in the sequence of shapes called accent tones. The uniqueness of a speaker or a task lies in the shapes that he employs - the length of the shape, the position of peak, the overall shape (amount of rise or fall or rise+fall etc.,). Intonational phonology postulates that there are only a finite number of shapes that characterize the F0 contour for a language, dialect or a linguistic type. This can be seen by the limited number of tones used in the ToBI annotation scheme [3]. Given this hypothesis, its then natural to talk about conversion between these shapes[1] of different speakers/styles of speech.

Most voice conversion frameworks use frame level representations of F0 (in the order of 5-10 milliseconds) that are ill-equipped to capture prosodic phenomena that are spread over longer ranges, that of syllables, metrical feet and beyond. Usually a variant of pitch range adaptation [4] is employed where the source F0 is transformed to the target speaker by employing the z-score transformation at the frame level –

$$F0_{(t)}^{tgt} = \frac{\sigma^{tgt}}{\sigma^{src}} \left(F0_{(t)}^{src} - \mu^{src} \right) + \mu^{tgt}$$

where $F0_{(t)}$ is the fundamental frequency at time instant t and μ, σ denote the mean and standard deviation of F0's from the training and adaptation data for the source and target speakers respectively. Therefore, while the pitch range is mapped to the target speaker, the transformation is blind to the aspects of identity and style that are spread over much larger contexts than the frame.

Here, we attempt a more informed conversion of a source speaker's accent tones to those of a target speaker. As a first step in modeling such a conversion, we need to decide the scope within which we want to analyze the accent tones of the two speakers. While syllables have been traditionally used in the past [5] for anchoring the accent tone and simulating parallel data, the style that we intend to capture in this work could be spread over multiple syllables, phonologically referred to as a stress group or metrical foot. Believing in the phonological hypothesis that the F0 is structured for conveying linguistic meaning of the underlying text, our goal is to convert an unseen F0 contour of the source speaker and predict the likely contour (with the appropriate accents) that the target speaker may have produced in delivering that sentence. For capturing speaking style, it makes practical sense to consider an accent in context of its neighboring syllables that exist within the same accent tone. In Sec 3, we propose a method to detect such regions, which we refer to as feet. The superpositional Statistical Phrase/Accent Model (SPAM) of intonation [6] is used to separate the phrase and accent regions that are parameterized as TILT vectors [7]. While the framework itself prescribes syllable-based accents, we use the method at the level of metrical feet, so that the discussion of conversion of accent tones across speakers is more tangible.

2. DATA AND PREPROCESSING

Since the goal here is to to learn a mapping between the contours within a phonological context, we choose the CMU ARCTIC

[1] Phonological frameworks are usually qualitative and a pitch accent/tone is only described but not parameterized. In this work, we treat it quantitatively.

databases [8] which has multiple male/female speakers delivering the same set of sentences, with roughly 1 hour of speech each.

The Pitch extraction was done using the *get_f0* utility of ESPS [9] for every 5 millisecond region of speech. While F0 exists only for purely voiced regions of the speech, listeners perceive as if there is an interpolation through the unvoiced regions [10]. We simulate this by applying smoothing and simple interpolations through unvoiced regions (except pauses) to better model intonation as a contour. It is to be noted that F0 extraction is still fraught with errors like pitch halving which can underestimate any attempt to model continuous regions of F0. The databases were automatically segmented using the *ehmm* utility of the festvox suite [11] and Festival utterance structures were built to enable linguistic representation in tandem with F0 analysis.

3. AUTOMATIC EXTRACTION OF METRICAL FEET

In order to model the correspondence and learn a transformation between the tone accents of two speakers, it is necessary to analyze the F0 contours within a phonologically identical context. As discussed before, in order to model multi-syllable prosodic phenomena, we analyze each accent at the level of the *metrical foot*, where a foot may be defined as consisting of an accented syllable, followed by all unaccented syllables that precede the next accented syllable or a phrase boundary. In principle, this modeling is similar to the one proposed in [12], however the data used in the latter is manually annotated and cleaned up. Since we are interested in analysis and conversion to/from arbitrary speaker pairs, we device an automatic method to detect foot like regions within the F0 contour and parameterize them.

To evaluate the effectiveness of the parameterization and the representational level, we compute reconstruction errors and correlations. An optimal representational scheme gives a small reconstruction error and preferably has some theoretical basis. In earlier work [6], it has been shown that for statistical parametric speech synthesis [13], syllable is more suited as a unit for intonation than the frame. [6] employs a superimpositional model comprising accents and phrases that are added in the $log(f0)$ domain. The phrase component is modelled as a decision tree at the segment level that models both long range phenomena like declination and segmental perturbations (microprosody). The residual accents are parameterized as syllable-linked TILT vectors [7]. In this work, we use the same model for the phrase components but the accent residuals are parameterized at the level of metrical feet.

To detect feet like regions, we run the constrained iterative component extraction method described in [6] to decompose all the $log(f0)$ contours into their globally optimal phrase and syllable level accent residuals. We then run another pass of analysis over all the syllables, re-analyzing groups of syllables (referred in the algorithm as a stress group, a having only a single primary stressed syllable in the group), so as to obtain the least reconstruction error. We also impose an additional constraint that a feet can end only at a word boundary. This is done due to practical concerns arising from feet ending within a word during voice conversion. The procedure is presented in Algorithm 1 and an illustration of the example reconstruction from the detected foot regions using automatically analyzed parameters is presented in Fig. 1

The output of the algorithm for each utterance is a sequence of TILT vectors, one per each foot that is detected on the contour. A TILT vector consists of 4 values (peak position, total length of the event, duration of the event and tilt, a quantified shape parameter). It also outputs the linguistic context, i.e., the words underlying each foot. These are easily synthesizable as is shown in the Fig 1. It can be seen that the foot based representation is quite good in that it can parameterize the salient peaks and the overall shape with minimal loss in error and correlation, and removes some unnecessary fluctuations within the contour that would be inevitable with a lower level representation. Table 1 presents the mean error and correlations for natural and resynthesized contours of the parameterized feet regions. It is rewarding that the same contour can be represented in about half or lesser number of parameters with minimal reconstruction loss, going from syllable to feet level representations, in addition to giving a simple parametrization of the speaker style.

Algorithm 1: Automatic Metrical Feet Extraction Method

```
for all Phrases do
    foot initialized
    predict $phrase$ contour using a statistical model
    $f0\_residual = f0 - phrase$
    for all Words do
        for all Syllables do
            add syllable to stress_group
            $syl\_accent = tilt\_analyze(f0\_residual)$ over syllable
            $syl\_error = f0\_residual - tilt\_resynth(syl\_accent)$
            $foot\_accent = tilt\_analyze(f0\_residual)$ over stress_group
            $foot\_error = f0\_residual - tilt\_resynth(foot\_accent)$
            if ( $foot\_error \geq prev\_foot\_error + syl\_error$) AND (word boundary found) then
                foot= stress_group - { current syllable}
                foot ended on previous syllable
                output $prev\_foot\_accent$
                stress_group = current syllable
                $prev\_foot\_error = syl\_error$
                $prev\_foot\_accent = syl\_accent$
            else
                $prev\_foot\_error = foot\_error$
                $prev\_foot\_accent = foot\_accent$
            end if
        end for
    end for
    if stress_group $\neq \phi$ then
        foot ends at Phrase boundary
        output $prev\_foot\_accent$
        stress_group $= \phi$
    end if
end for
```

For the current purpose of voice conversion, the main idea is to model any systematic way in which the nuclear accent(the peak) moves about from the source to the target and how the shape of the accent transforms over the feet. So, the algorithm given here is run on the source speaker and the feet are detected. It should be noted that different speakers may have a different set of metrical feet they choose to employ. For the arctic speakers, on an average, there is only about 38% of the feet matching per utterance for a random speaker pair. So, it is not easy to get parallel data with this setting. We deal with this problem by ‘force aligning’ the source speaker’s feet on the target speaker, so an analysis is carried out on target speaker’s speech within the same linguistic context, so that there is an alignment of the number of feet analyzed in each utter-

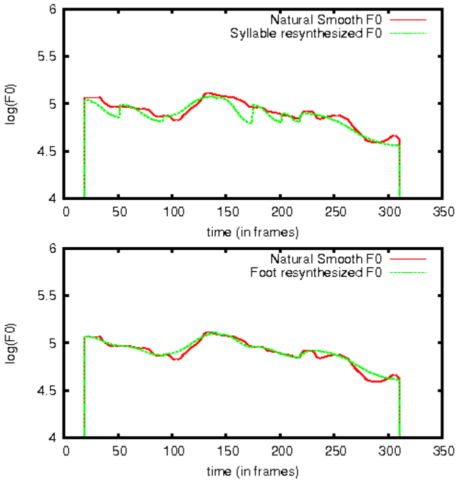

Fig. 1. Illustrating syllable vs foot based reconstruction using the algorithm described above

Table 1. Errors and correlations on resynthesized contours using different representational levels

Speaker	Syllable		Foot	
label	RMSE	CORR	RMSE	CORR
awb	0.13	0.77	0.14	0.73
bdl	0.09	0.82	0.12	0.76
ksp	0.08	0.79	0.11	0.73
rms	0.10	0.80	0.14	0.72
slt	0.07	0.73	0.09	0.69

ance pair. The contours over each feet are then parameterized using the TILT representation, which stores the peak, the total length over the contour, duration and the tilt shape parameter of the accent for both the source and target speakers. The corresponding TILT parameterized vectors of each foot form the parallel data, from which to model a transformation. Table 2 shows the correlation matrix for the shape parameter(tilt) of corresponding feet in each speaker pair. Note that this matrix is not symmetric because the feet boundaries vary as a different speaker is chosen as the source. It still is satisfying that there is a small but positive correlation between almost all speaker pairs on the shape parameter.

Table 2. Correlation matrix for the `tilt` shape parameter among speakers for corresponding feet

	awb	bdl	ksp	rms	slt
awb	1	0.139	0.333	0.293	0.254
bdl	0.155	1	0.290	0.244	0.250
ksp	0.336	0.218	1	0.301	0.260
rms	0.256	0.202	-0.01	1	0.213
slt	0.230	0.162	0.137	0.158	1

4. PROPOSED 2-LEVEL F0 CONVERSION TECHNIQUE

The SPAM intonation model represents $log(f0)$ as a sum of two components, the phrase, that models the long term trend of the contour and accents that model the local detours. In this work, we use the simple z-score transform as in Equation 1 on the phrase contour because the mean pitch is roughly determined in the phrases and a simple affine transformation is sufficient to approximate the target speaker's phrase components. Accents are however complex since they are described by many parameters that bear information about the speaker style. To transform accents, we train a mapping between the two speakers' accent vectors using parallel data as described below. An illustration of the proposed 2 level F0 conversion is shown in Fig 2.

The goal of the mapping that we learn from the parallel data of accent vectors is to apply it to accent shapes of an unseen utterance of the source speaker, and predict corresponding shapes of the target speaker. To accomplish this, we use the Gaussian mixture model(GMM) based technique often used for spectral conversion [14]. The conversion can be realized by a continuous mapping based on soft clustering of the parallel accent features [15].

Let x_t and y_t be the TILT accent vectors for each metrical foot. The joint probability density of the source and target vectors is modelled as the following GMM -

$$P(z_t|\lambda^{(z)}) = \sum_{m=1}^{M} w_m \mathcal{N}(z_t;\mu_m^{(z)},\Sigma_m^{(z)})$$

where z_t is the joint vector $\begin{bmatrix} x_t' \\ y_t' \end{bmatrix}$, with the GMM having M mixtures with a mean, covariance and mixture weight of the m'th Gaussian component denoted by w_m, $\mu_m^{(z)}$ and $\Sigma_m^{(z)}$ respectively. The Covariance matrix $\Sigma_m^{(z)}$ is constrained to be of the form $\Sigma_m^{(z)} = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix}$, where each partial covariance matrix is set to be a full matrix, because some TILT parameters (duration and tilt amplitude) have positive correlation between themselves [7].

At test time, given an accent shape x_t of the source speaker, we want to predict the corresponding y_t of the target speaker as follows –

$$\begin{aligned} \hat{y}_t &= \sum_{i=1}^{M} p(m_i|x(t),\lambda^{(z)})E(y_t|x_t,m_i,\lambda^{(z)}), \\ E(y_t|x_t,m_i,\lambda^{(z)}) &= \mu_i^{(y)} + \Sigma_i^{(yx)}\Sigma_i^{(xx)^{-1}}(x_t-\mu_i^{(x)}), \\ p(m_i|x(t),\lambda^{(z)}) &= \frac{w_i\mathcal{N}(x_t;\mu^i,\Sigma_i^{(xx)})}{\sum_{j=1}^{M} w_j\mathcal{N}(x_t;\mu_j^{(x)},\Sigma_j(xx))} \end{aligned}$$

5. EXPERIMENTS AND RESULTS

To evaluate the proposed transformation, we select speakers `awb, ksp, slt and rms` of the Arctic databases. SPAM intonation models were trained on the training data (90%) for each speaker. For each selected speaker pair, a transformation data of 200 sentences (about 12 minutes of speech) is randomly selected. The source speaker's intonation is analyzed as described in Sec 3. Since the transformation data is relatively small, a phrase CART tree cannot be trained for the target speaker, so the phrase model of the source speaker is used on the target speaker's utterance to predict a possible phrase contour. The phrase contour is shifted along the $log(f0)$ axis such that the residuals are all non-negative with a minimum at 0. For each metrical foot, the accent residual of the target speaker is also analyzed within the same linguistic context, to obtain a parallel set of accents for the speaker pair. GMM Joint densities are trained

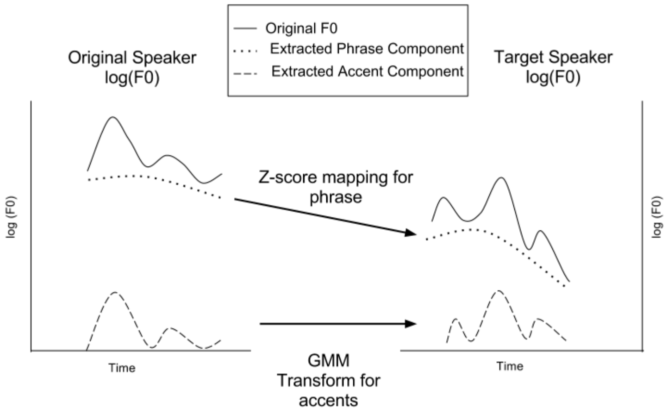
Original F0
Extracted Phrase Component
Extracted Accent Component
Original Speaker log(F0)
Target Speaker log(F0)
Z-score mapping for phrase
log (F0)
log (F0)
Time
Time
GMM Transform for accents

Fig. 2. Schematic diagram of proposed F0 conversion technique for phrase and accent components

and the mapping function described in Sec. 4 is computed. Also the means and standard deviations of the phrase components are computed to learn a z-score transformation for the phrase components of the two speakers.

For a test set of 100 sentences, the source speech is analyzed, the feet extracted and parameterized. The durations are modified in the parameterization to match those of the reference speech of the target speaker. This is done so as to be able to objectively compute the root mean squared error (`rmse`) and correlation (`corr`) metrics for each utterance. The SPAM phrase model of the source speaker is used to predict a phrase curve over the source speaker, and the phrase level z-score transform is applied to predict the appropriate phrase contour for the target speaker. GMM transform is applied on the TILT parameterizations of the accents over each foot, to predict the possible accent shapes of the target speaker. Resynthesis of the transformed parameters is done and added with the transformed phrase contour to predict the F0 contour for the target speaker for the durations he employed. As a baseline to compare against, we use the traditional z-score mapping directly on the $log(f0)$ contour – the resynthesized parameters of the source speaker for the durations of the target and the result mapped to the mean and range of the target speakers $log(f0)$.

The predicted contours of the baseline and the proposed approaches are evaluated against the reference target speaker $log(f0)$s, using rmse and correlation measures. Table 3 shows the average measures over the test set for several speaker pairs. All statistically significant differences in correlation are shown in bold font. It can be seen that the correlation of the transformed contours of the proposed approach are consistently improved compared to the baseline z-score mapping on the F0 contour. Large scale listening tests are currently being done using the Amazon Mechanical Turk.

6. CONCLUSIONS

In this paper, we propose an approach for F0 transformation for voice conversion. An automatic method is designed to extract metrical feet within F0 contours. Corresponding feet of two speakers speaking the same underlying text are extracted and parameterized using 4-dimensional TILT vectors. A Gaussian mixture model based mapping is trained between the foot-based accent vectors. This mapping is used to convert unseen contours of utterances of the source speaker to predict the likely contour of the target speaker. Objective evaluations prove that the method is significantly better than the baseline frame level, z-score mapping technique.

Table 3. Objective comparison of frame level z-score transformation and GMM transformation of feet based accent vectors

Speaker	Z-score transform		Foot based	
pair	RMSE	CORR	RMSE	CORR
`bdl-slt`	0.494	0.377	0.466	**0.521**
`bdl-ksp`	0.264	0.450	0.289	**0.526**
`bdl-awb`	0.305	0.528	0.310	**0.647**
`bdl-rms`	0.593	**0.461**	0.421	0.405
`ksp-bdl`	0.324	0.557	0.312	0.556
`ksp-slt`	0.470	0.423	0.438	**0.505**
`ksp-rms`	0.493	0.339	0.697	**0.513**
`ksp-awb`	0.334	0.561	0.304	**0.631**
`rms-bdl`	0.216	0.565	0.238	**0.590**
`rms-slt`	0.628	0.247	0.443	**0.487**
`slt-bdl`	0.638	0.465	0.350	**0.491**
`slt-rms`	0.915	**0.531**	0.475	0.307

7. ACKNOWLEDGEMENTS

Support for this research was provided by the Fundação para a Ciência e a Tecnologia through the CarnegieMellon|Portugal program under grant SFRH/BD/33510/2008, and through projects PEst-OE/EEI/LA0021/2011, CMU-PT/HuMach/0039/2008 and CMU-PT/0005/2007.

8. REFERENCES

[1] E. Zetterholm, "Same speaker different voices: A study of one impersonator and some of his different imitations," in *Intl Conf on Speech Science and Technology*, 2006, pp. 70–75.

[2] Y. Stylianou, "Voice transformation: A survey," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 3585 –3588.

[3] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: a standard for labelling English prosody.," in *Proceedings of ICSLP92*, 1992, vol. 2, pp. 867–870.

[4] T. Toda, A.W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222 –2235, nov. 2007.

[5] E. Helander and Janu Nurminen, "A Novel method for prosody prediction in voice conversion," in *ICASSP 2007*, Hawaii, 2007.

[6] Gopala Krishna Anumanchipalli, Luis C. Oliveira, and Alan W Black, "A Statistical Phrase/Accent Model for Intonation Modeling," in *Interspeech 2011*, Florence, Italy, 2011.

[7] P Taylor, "Analysis and synthesis of intonation using the tilt model," *Journal of the Acoustical Society of America*, vol. 107 3, pp. 1697–1714, 2000.

[8] J. Kominek and A. Black, "The CMU ARCTIC speech databases for speech synthesis research," Tech. Rep. CMU-LTI-03-177 http://festvox.org/cmu_arctic/, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 2003.

[9] David Talkin, “ESPS, Entropic Research Lab Inc.,,” 1993.

[10] Paul Taylor, *Text-to-Speech Synthesis*, Cambridge University Press, 2009.

[11] A. Black and K. Lenzo, “Building voices in the Festival speech synthesis system,” http://festvox.org/bsv/, 2000.

[12] Esther Klabbers and J.P.H. van Santen, “Clustering of foot-based pitch contours in expressive speech synthesis,” in *ISCA Speech Synthesis Workshop V*, Pittsburgh, PA, 2006.

[13] Alan W Black, “Clustergen: A statistical parametric synthesizer using trajectory modeling,” in *Interspeech 2006*, Pittsburgh, PA, 2006.

[14] Y. Stylianou, O. Cappé, and E. Moulines, “Statistical methods for voice quality transformation,” in *Eurospeech95*, Madrid, Spain, 1995, pp. 447–450.

[15] A. Kain, *High Resolution Voice Transformation*, Ph.D. thesis, OGI School of Science and Engineering, Oregon Health and Science University, 2001.