SPEAKER INTONATION ADAPTATION FOR TRANSFORMING TEXT-TO-SPEECH SYNTHESIS SPEAKER IDENTITY

Mahsa Sadat Elyasi Langarani, and Jan van Santen

Center for Spoken Language Understanding, Oregon Health & Science University, Portland, OR, USA

{elyasila, vansantj}@ohsu.edu

ABSTRACT

In this study, we propose a new intonation adaptation method to transform the perceived identity of a Text-To-Speech system to that of a target speaker with a small amount of training data. In the proposed method, during training we fit parametrized accent and phrase curves to parallel recordings of the target speaker F_0 curves, and estimate the parameters of a mapping between the corresponding parameter spaces. During test, we fit the accent and phrase curves to the source utterances, apply the mapping, and create an F_0 contour from the mapped accent and phrase curves. We compare the proposed method with a baseline adaptation method in which the source F_0 contour is transformed linearly such that the per-utterance mean and variance of the target F_0 contour is left unaltered. Perceptual tests showed that the proposed method was better than the baseline method in two subjective tests that assess similarity to the target speaker and speech quality, respectively.

Index Terms: Prosody, Intonation modeling, Text-to-Speech synthesis, Adaptation

1. INTRODUCTION

Speaker identity plays an important role in human-human and human-computer communication. Computer-generated speech intended to sound like a specific *target speaker* generally uses spectral feature mapping. In the case of voice conversion (VC) in which any utterance spoken by a *source speaker* is converted to an utterance that sounds as if spoken by the target speaker, possibilities are limited because the contents is not known. Specifically, mimicking the dynamic details of the target speaker's prosodic style is difficult with spectral mapping. Yet, prosodic style is an extremely important aspect of speaker identity [1, 2]. However, in Text-To-Speech synthesis (TTS), the system has full access to textual contents, phonemes, and temporal information. In this report, we will show how this information can be used for speaker-specific F_0 generation using limited training data. To clarify, in the case of TTS, the source speaker is the speaker whose recordings were used to generate the acoustic units (for unit selection approaches), acoustic inventory (for diphone based synthesis), Hidden Markov Model (HMM) parameters (for HMM based synthesis), or whatever speech data are used during synthesis. This speaker's recordings may also be used as training data for prosody mimic. Thus, the speech generated by a TTS system generally sounds like the source speaker. For prosody mimic, the challenge is to compute a transformation that, when applied to the speech data or to any representations thereof, generates output speech mimicking a target speaker.

This material is based upon work supported by the National Science Foundation under Grant No. 0964468.

We note that F_0 can be analyzed at different levels (last column in Table 1: frame [3], phone [4], syllable [5, 6, 7, 8, 9], phrase [7, 9], and sentence [4]). A fundamental issue is which level is most appropriate for capturing a key property of F_0 movement, which is that — except where perturbed or interrupted by obstruents — it has a *smooth polysyllabic shape* with typically *no more than two inflection points*. For example, in English, standard H*L intonation accents involve a smooth rise in course of the accented syllable followed by a descent until the next accented syllable or phrase boundary [10, 11, 12, 13]. A recent study explicitly addressing this issue [14] considered various phonological units in a statistical parametric speech synthesis framework, including the frame, syllable, word, accent group, phrase, and sentence. "Accent group" was defined as a sequence of syllables containing an accented syllable and not necessarily as a (left-headed) foot, which requires that the first syllable is accented (e.g., [15, 16, 2]). Anumanchipalli [14] showed that the best-performing phonological unit is the accent group. This result suggests that we may need to consider units that are larger than the syllable and that, in addition, do not need to coincide with word boundaries.

As indicated in the third column in Table 1, various levels have been used to represent F_0. Since most TTS adaptation and VC approaches are focused on spectral features, they use simple methods to capture prosody [17, 18, 19]. Typically, F_0 is represented by just the mean and the standard deviation (SD); thus, during synthesis, the output utterance will match only these target speaker features without attempting to capture the dynamic details of the speaker's prosodic style [4]. In a more sophisticated approach, Chappell proposed a scatterplot model of the mean of F_0 with one data point per voiced phone to model F_0 [4]. Patterson went a step beyond Chappell's approach and used four data point (sentence-initial and final F_0 values, accent peaks and valleys) in an utterance to represent F_0 [20]. HMM-based [8, 3] and superpositional [13, 21, 22, 9] approaches are potentially yet more accurate and practical methods for capturing intonation. In [23], we showed that a *data-driven* foot-based intonation generator method ("DRIFT") produces more natural sounding F_0 contours than a standard HMM based method. DRIFT employs a model-based F_0 generation method that guarantees that contours will have a smooth polysyllabic shape [22, 24]. In contrast to Anumanchipalli et al. [14, 25], the phonological unit used in DRIFT is the foot. In the current study, we show how we can use DRIFT for F_0 mimic.

Intonation can be transformed at different levels (second column in Table 1): frame [26, 4, 27, 3], tone [28] syllable [6, 5, 7, 29, 8, 28], word [30], sequence of syllables [29, 8, 31] and sentence [4] with different methods(first column in Table 1). As mentioned, the most common method for transform F_0 is by globally matching the mean

Approach	Adaptation method	Adaptation domain	Intonation model	model domain
baseline	Linear	Frame-level	Mean and SD of raw intonation	Sentence-level
[4]	Linear	Frame-level	Average(mean and SD) of raw intonation	Sentence-level
[4]	Polynomial conversion	Frame-level	Scatterplot model of mean intonation	Phone-level
[26]	Piecewise linear mapping	Frame-level	Pitch range model	Accent-level and sentence-level
[7]	Linear modification	Syllable-level	Raw intonation(professional recorded)	Syllable-level and phrase-level
[7]	GMM	Syllable-level	Pitch target model	Syllable-level
[9]	GMM	Syllable-level	DCT+Multi-level dynamic features	Syllable-level and phrase-level
[8]	Data-driven F0 segment selection	Sequence of syllable	MSD-HMM	Syllable-level
[7]	CART	Syllable-level	Pitch target model	Syllable-level
[5]	Codebook + CART	Syllable-level	DCT	Syllable-level
[4]	Contour codebook + DTW	Sentence-level	Raw intonation	Sentence-level
[3]	MSD-MLLR	Frame-level	MSD-HMM	Frame-level
[6]	MLLR	Syllable-level	GMM	Syllable-level
Proposed method	JDGMM	Foot-level and phrase-level	DRIFT	Foot-level and phrase-level

Table 1. A comparison among few approaches for intonation transformation. Approaches are classified according to four categories: Adaptation method, adaptation domain, intonation model, and model domain.

and SD of the target speaker's F_0 contour. The means and SD's of the source and target speaker's F_0 contours define a linear transformation that is applied to the source speaker's F_0 contour, typically in the log domain [4]. Extensions of this approach include higher-order polynomial [4], piecewise linear transformation [26] and linear modification based on hand-labeled intonational (syllable-phrase) features. Another class of methods predict intonation by modeling F_0 and spectral features jointly [32, 33, 34]. Due to the limited amount of data, statistical techniques are usually utilized in extracting the mapping function. The most popular technique is based on a Gaussian mixture model (GMM) [7, 29, 27, 9, 35]. Two other methods are studied by F_0 contour codebook [4] and parametrized codebook [5, 29]. Weighting multiple contours has shown performance improvement [36]. Various other methods such as hierarchical models [37], CART [5, 7] and MLLR [6, 3] are proposed. A comparison of some of the mentioned approaches is studied [38].

In the present study, we propose a new intonation adaptation method to transform the perceived identity of a TTS system to that of a target speaker with a small amount of training data. For modeling intonation, we employ our DRIFT method that captures F_0 with a small number of parameters at two levels: the foot and the phrase. Because the number of parameters to be estimated is relatively small, it is feasible to adapt speaking style using any mapper function such as Joint distribution Gaussian mixture model (JDGMM). We compare the proposed method with a baseline adaptation method in which the source F_0 contour is transformed linearly such that the per-utterance mean and variance of the target F_0 contour is unaltered; yet, this generated F_0 contour still has the dynamics of the source F_0 contour. Thus, this study asks two questions. First, is mimicking just the mean and SD enough? And, if not, does our method succeed in capturing this extra, dynamic information that is lost in the linear transformation approach? We designed two subjective listening experiment (speech similarity and speech quality) to study performance of the two methods, for two male target speakers.

2. DATA-DRIVEN FOOT-BASED INTONATION GENERATOR (DRIFT)

2.1. Intonation model

In a previous study [22], we proposed a new method for decomposing a continuous F_0 contour — interpolated in unvoiced regions — into component curves in accordance with the *General Superpositional Model* (GSM; [13]), a broad generalization of the Fujisaki model [21]: a phrase curve ($P(t)$ in Equation 1) and a sum of one or more accent curves ($A(t)$ in Equation 1).

$$F_0(t) = P(t) + A(t) \tag{1}$$

In this method, the phrase curve consists of two log-linear curves, between the phrase start and the start of the phrase-final foot (generally associated with the nuclear intonation accent), and between the latter and the end point of the last voiced segment of the phrase, respectively. We use a combination of the skewed normal distribution and a sigmoid function to model three different types of accent curves. First, the skewed normal distribution is employed to model rise-fall accents that occur in non-phrase-final positions as well as, in statements, in utterance-final positions (f in Equation 2). Second, a sigmoid function is used to model the rise at the end of a yes/no question utterance (g in Equation 2). And, third, the sum of the skewed normal distribution (f) and the sigmoid function (g) is used to model continuation accents at the end of a non-utterance-final phrase (h in Equation 2). The number of accent curves (n) is equal to the number of feet in a phrase (equation 2). The a and b parameters are binary and are used to compactly express the three accent types as

$$A(t) = \sum_{i=1}^{n} (b_i(a_i f(t) + (1 - a_i)g(t)) + (1 - b_i)h(t)). \tag{2}$$

For example, a yes/no question sentence with two feet (rise-fall ($\%LH^*L$) and yes/no question ($L^*H\%$) accent types) is represented by $a_1 = 1$, $b_1 = 1$ and $a_2 = 0$, $b_2 = 1$, respectively. In Equation 3 and 4, C and D stand for the amplitudes of the accent curves. The two parameter sets $\{\omega, \xi, \alpha\}$ and $\{\beta, \gamma\}$ indicate {scale, location, skewness} of the skewed normal distribution, and {slope, location} of the sigmoid function. These parameters together with the three parameters of the phrase curve are optimized using Sequential Least Squares Programming (for details, see [22]).

$$f(t) = C\frac{2}{\omega}\phi(\frac{t-\xi}{\omega})\Phi(\alpha(\frac{t-\xi}{\omega})) \tag{3}$$

$$g(t) = D\,\frac{1}{1+e^{-\beta(t-\gamma)}} \tag{4}$$

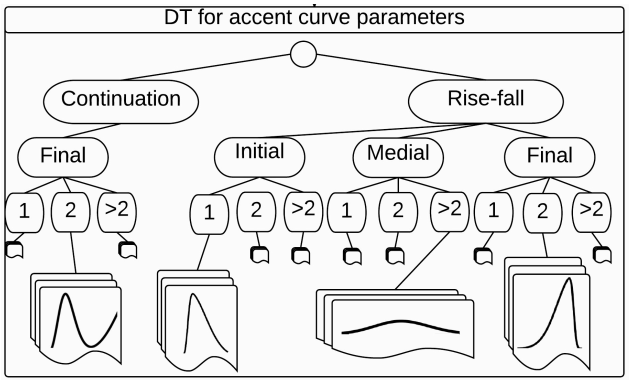


Fig. 1. Decision-tree for accent curves parameters

2.2. Analysis

In order to segment training utterances (subsection 5.1) into foot sequences, our method uses three contextual features: accent labels, syllable labels, and phrase boundaries, to automatically create foot boundaries.We extract five contextual features per foot:

$$Set_{Acc} = \begin{cases} PT\text{: phrase type (statement, continuation)} \\ FPos\text{: foot position in phrase (initial, final, other)} \\ SNum\text{: number of syllables in foot (1, 2, >2)} \\ OD\text{: onset duration of stressed accend syllabel} \\ RD\text{: rime duration of stressed accend syllabel} \end{cases}$$

An *accent curve inventory* is created as follows. For each training utterance, we extract F_0 and then fit the intonation model described in subsection 2.1 to compute the phrase curve and the accent curve parameters. We store the vector comprising the estimated accent curve parameters and the values of OD and RD in the inventory. The inventory contains twelve *sub-inventories* defined in terms of the Set_{Acc} features PT, $FPos$, and $SNum$ (Figure 1). (Because the data were not tagged for y/n (or any) questions, we did not include a y/n question sub-inventory.)

A *phrase curve inventory* is created as follows. We extract two contextual features per phrase:

$$Set_{Phr} = \begin{cases} PT\text{: phrase type (statement, continuation)} \\ FNum\text{: number of feet in phrase (1, >1)} \end{cases}$$

After extracting the phrase curve and the accent curve parameters for each training utterance using the intonation model described in subsection 2.1, We store a vector consisting of the estimated phrase curve parameters (phrase start, the start of the phrase-final foot, and phrase final) in the inventory. Note that if a phrase contains just one foot, then the phrase is modeled by two parameters (phrase start and phrase final). The inventory contains four *sub-inventories,* differentiated in terms of the Set_{Phr} features, PT *and* $FNum$.

2.3. Synthesis

In the proposed method, an input sentence is segmented into phrases, each phrase is segmented into a foot sequence, and for each foot the Set_{Acc} features are extracted. The first four features are extracted from text data, and the values of OD and RD are predicted using force alignment applied on original utterances [39]. A suitable accent sub-inventory is chosen for that foot by traversing the proposed decision tree using the first three features: PT, $FPos$, and $SNum$ (Figure 1). We calculate the euclidean distance between the OD, and RD of the current foot and the stored accent curves in the chosen sub-inventory. The five candidate accent curves with the lowest distance in that sub-inventory are retrieved. To minimize the differences between successive accent curve heights in a phrase, we apply a Viterbi search to the sequence of candidate accent curves; the observation matrix consists of the normalized duration distances and the transition matrix consists of the normalized accent curve height differences.

For the current phrase, the suitable phrase sub-inventory is chosen by using these two features: PT and $FNum$. We use the average of the stored phrase curves parameters in the chosen sub-inventory as synthetic phrase curve parameters.

3. INTONATION MAPPING

3.1. Baseline Mean-Variance Linear mapper

In VC and TTS literature, it is often assumed that the F_0 mean and SD are adequate to capture prosodic style [40]. The most common method for transforming F_0 is to globally match the average mean and average SD of the target speaker's F_0 contour, while maintaining the dynamic intonation pattern of the source. With this assumption, intonation can be transformed by mapping log-F_0 using a linear transformation, where μ and σ represent average mean and SD of the log-F_0 of the training set [4].

$$F_{mimicked} = \frac{\sigma_{target}}{\sigma_{source}} \left(F_{source} - \mu_{source}\right) + \mu_{target} \qquad (5)$$

For the baseline method, we use a slightly different linear transformation, in which the baseline does not have a training stage. Therefore in our baseline method, μ and σ represent the mean and SD of the original utterances of the *test* set. This assumption gives the linear model a strong opportunity to over-fit the target speaking style in a given sentence, making it in principle more effective than the average-mean-and-SD linear mapper.

3.2. Joint distribution GMM mapper

In this section, we briefly overview the GMM mapping function [41]. Let $X = \{x_1, ..., x_n\}$ and $Y = \{y_1, ..., y_n\}$ be set of parameters vector for n segments (foot or phrase in case of mapping accent parameters or phrase parameters, respectively) from the source and target model, respectively. Note that each vector is normalized using maximum and minimum of X and Y. Let $Z = [X, Y]$ is the joint source-target parameters vector. A GMM represents the distribution using M multivariate Gaussian

$$P(z) = \sum_{m=1}^{M} \alpha N(z; \mu_m, \Sigma_m) \qquad (6)$$

where $N(z; \mu_m, \Sigma_m)$ is a normal distribution with mean μ_m and covariance Σ_m of component m. Prior probability of the component m is represented by α_m. The parameters of the GMM are calculated using the Expectation Maximization (EM) algorithm on the joint vector Z. During transformation, for each component, we estimate the weighted mixture of maximum likelihood estimator of the target vector given the source vector for each component

$$\hat{y}_i(x_i) = E[Y|X = x_i] = \sum_{m=1}^{M} w_m^x(x_i)\left[\mu_m^y - \Sigma_m^{xy}\Sigma_m^{xx-1}(x_i - \mu_m^x)\right] \quad (7)$$

where $w_m^x(x_i)$ is a posterior probability that the segment x_i belongs to the class described by the component m.

$$w_m^x(x_i) = \frac{\alpha_m N(x_i; \mu_m^x, \Sigma_m^{xx})}{\sum_{k=1}^{M} \alpha_k N(x_i; \mu_k^x, \Sigma_k^{xx})} \quad (8)$$

4. INTONATION ADAPTATION

4.1. Training procedure

The aim of F_0 adaptation is to predict the intonation style of the target speaker with a small amount of parallel training data, since otherwise one might just as well obtain a complete set of speech recordings of the target speaker and avoid the transformation process altogether. We randomly select a small set of recordings (subsection 5.1, 28 parallel utterances) from the source and target speakers. For each utterance, we apply the intonation model (described in subsection 2.1) to decompose the F_0 contour of the utterance into accent and phrase curves. We use the estimated source and target accent curve parameters to train a JDGMM mapper with two components ($M = 2$). This process is performed similarly done for phrase curve parameters. Thus, the mapper operates in the parameter space defined by the DRIFT model and *indirectly* maps source F_0 contours onto target F_0 contours (Top block-diagram in Figure 2_a).

4.2. Adaptation procedure

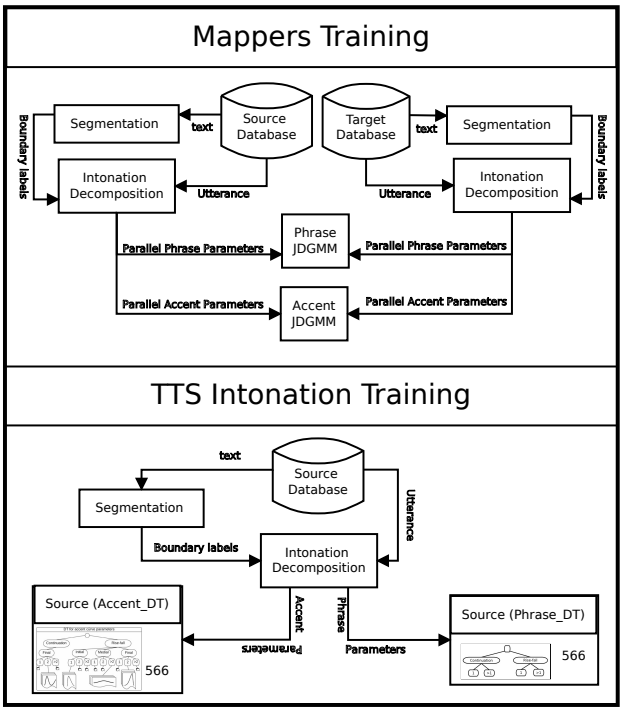
In this study, an input sentence is segmented into phrases, each phrase is segmented into a foot sequence, and for each foot the Set_{Acc} features are extracted. The first four features are extracted from text data, and the value of OD and RD are predicted using force alignment applied to the original utterance [39]. The five candidate source accent curves with the lowest distance in the selected sub-inventory are retrieved (see subsection 2.3). By applying the accent mapper to each five candidates, five transformed accent curves are predict per foot. To minimize the differences between successive transformed accent curve heights in a phrase, we apply a Viterbi search to the sequence of the transformed accent curves; the observation matrix consists of the normalized duration distances and the transition matrix consists of the normalized transformed accent curve height differences.

For the current phrase, the Set_{Phr} features are extracted. Parameters of the source phrase are predicted by calculating the average of the stored phrase curves parameters in the selected sub-inventory. Transformed phrase parameters are estimated by applying the phrase mapper to predicted source phrase parameters. (Figure 2_b)

4.3. Synthesis procedure

During synthesis the mapper is applied to the source speaker's DRIFT model parameters (i.e., the parameters that would be used to generate TTS output during normal operation, (Bottom block-diagram in Figure 2_a)) to generate predicted target speaker DRIFT parameters (described in subsection 4.2); these predicted parameters are used to generate the accent and phrase curves, which are added together in accordance with the GSM to generate a target F_0 contour; finally, this target contour is used in the process of generating output speech.

a) Training



b) Adaptation

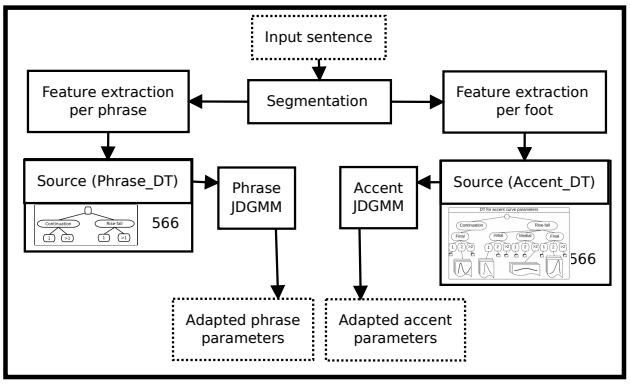


Fig. 2. Block-diagrams of training and adaptation of proposed method

5. EXPERIMENTS

5.1. Databases

For the TTS adaptation experiment, we use the CMU Arctic database [42]. We consider one professional US English female speaker (SLT) as source speaker and two male speakers (English speaker: BDL, and Scottish speaker: AWB) as target speakers.

This corpus contains 1132 utterances per speaker (parallel sentences), which are recorded at 16bit 32KHz. Utterances of SLT and BDL were recorded in a sound proof room while AWB's utterances were recorded in quiet office. The database is automatically labelled by CMU Sphinx using FestVox labeling scripts. No hand corrections are made.

We used two training sets for subjective evaluation: a large set, which included 566 training utterances, and a small set, which included 28 (5% of the large set) training utterances. We used the large set for training of the source model and the small set for training the mapper. A set of 150 utterances was selected randomly for test purposes.

5.2. Evaluation

For subjective evaluation of the intonation generation performance of the two approaches, we designed two tests: the first test measures speech quality and the second test measures speech similarity between stimuli and the target speaker. We used Amazon Mechanical Turk [43], with participants who have approval ratings of at least 90% and were located in the United States.

In each test, we evaluated the two approaches by imposing the F_0 contours generated by the two approaches onto recorded natural speech, thereby ensuring that the comparison strictly focused on the quality of the F_0 contours and was not affected by other aspects of the synthesis process. To ensure that the F_0 contours were properly aligned with the phonetic segment boundaries of the natural utterance, the contours were time warped so that the predicted phonetic segment boundaries corresponded to the segment boundaries of the natural utterance. To compute the segment boundaries of the natural utterance, we used the phoneme durations predicted by force alignment [39]. Finally, we used PSOLA to impose the synthetic contour onto the natural recordings[1].

5.2.1. *Speech quality test*

We used a comparison test to evaluate the quality of the F_0 contours synthesized by the two approaches. In this test, listeners hear two stimuli with the same content back-to-back and then are asked which they prefer using a five-point scale consisting of -2 (definitely first), -1 (probability first), 0 (unsure), +1 (probability second), +2 (definitely second) [44]. We randomly switched the order of the two stimuli. The experiment was administered to 150 listeners, with each listener judging 50 utterance pairs. Three trivial-to-judge utterance pairs were added to filter out unreliable listeners.

Figure 3-a shows the results for the test sets for two target speakers. For significance testing, we first compute a score for each utterance using Equation 9, and then, separately for each test set, apply a one-sample t-test. In Equation 9, j, n, m, and C_{ji} stand for j^{th} utterance of current test set, number of listeners, number of utterance of current test set, and the rating of the i^{th} listener for the j^{th} utterance, respectively, and $||$ indicates the absolute value.

$$score_j = \frac{\sum_{i=1}^{n}(C_{ji}|C_{ji}|)}{\sum_{j=1}^{m}(\sum_{i=1}^{n}(|C_{ji}|))} \quad , C_{ji} \in \{-2,-1,0,1,2\} \qquad (9)$$

Conventional *t*-test results showed that the scores of the two methods differed significantly from each other for AWB (First row of Table 2). We also performed a randomization test for the same difference by 2000 times randomly changing the signs of all ratings, computing the scores for each utterance, and calculating the *t* statistic. (This randomization test is more conservative than the conventional t-test.) The means and standard deviations of the resulting distributions are summarized in Table 2, and yield conclusions similar to those based on the conventional *t*-tests.

[1]The synthetic waves are available under following repository: http://cslu.ohsu.edu/~elyasila/wav_ASRU/

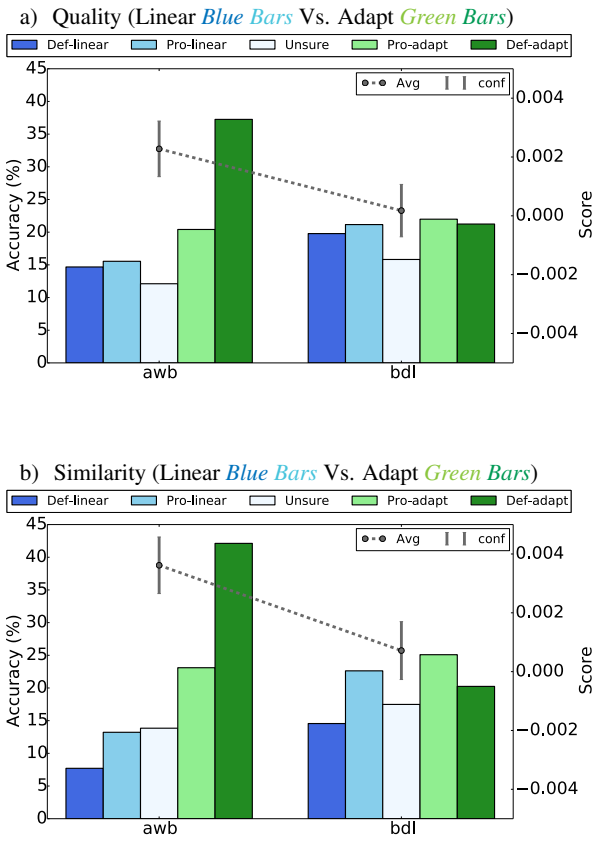


Fig. 3. Speech quality and similarity test. Dashed curves correspond to the values computed via Equation 9.

5.2.2. *Speech similarity test*

To evaluate speaker mimic accuracy, we designed a speaker similarity test. In this test, listeners heard three stimuli. First, a natural recording of the target speaker to convey the target speaking style. Second, two stimuli with the same content (but contents differing from that of the natural recording) back-to-back. They were then asked which provides the best mimic of the target using the same five-point scale as in the quality test (subsection 5.2.1). We randomly switched the order of the two stimuli. The experiment was administered to 150 listeners, with each listener judging 50 utterance pairs. Three trivial-to-judge utterance pairs were added to the experiment to filter out unreliable listeners.

Figure 3-b shows the results for the test sets for two target speakers. For speaker BDL, the baseline worked as well as the proposed method. This suggest that both the source (SLT) and target (BDL) have similar intonation patterns: matching the mean and SD appeared sufficient. However, for speaker AWB the proposed method

	AWB				BDL			
	t-test		Randomization		t-test		Randomization	
	t(149)	P-value	Mean	SD	t(149)	P-value	Mean	SD
L vs. **A (Quality)**	5.7749	1.0341e-08	1.5082	2.0859	0.4874	0.6264	0.6518	0.6406
L vs. **A (Similarity)**	8.8257	93139e-17	2.0077	3.2153	1.9756	0.0491	0.7022	1.0415

Table 2. Quality and similarity experiment results: one-sample t-tests [t-value(df), p-value], and mean and standard deviation (SD) of the randomization-based t-statistic distribution comparing the linear (L) and Adapt (A) methods, for two speakers (AWB and BDL)

		AWB					BDL				
		t(149)	P-value	Mean			t(149)	P-value	Mean		
				L	N	A			L	N	A
Mean of F_0	**L** vs. **N**	-0.5502	0.5826	149.0506	150.2694	–	-0.1684	0.8664	126.3289	126.5171	–
	A vs. **N**	-11.4206	2.2444e-25	–	150.2694	131.3975	-10.7368	5.1680e-23	–	126.5171	117.3713
SD of F_0	**L** vs. **N**	-2.5262	0.0121	47.2186	57.5239	–	-1.2474	0.2132	21.0202	23.0289	–
	A vs. **N**	-13.3240	3.2545e-32	–	57.5239	17.7554	-7.9373	4.0140e-14	–	23.0289	11.7965

Table 3. Differences in mean and SD between transformation methods and natural target speech: one-sample t-tests [t-value(df), p-value], and mean of the mean and standard deviation (SD) of F_0 of two speakers (AWB and BDL) for two pairwise comparisons of linear (L) and Adapt (A) methods with Natural (N) method.

was clearly superior , and marginally superior for BDL (Second row of Table 2).

Interestingly, for both speakers, the proposed method produced means and SDs that differed far more from those of the target speaker than the linear method 3. Yet, for both speakers , the proposed method was perceived as producing a significantly better mimic. Apparently, *copying the mean and SD of a target speaker is neither sufficient nor necessary for prosody mimic*.

6. CONCLUSION

The proposed method shows promise as a way to capture the dynamics of the F_0 contours of a target speaker. Whether it performs better than a much simpler linear transformation of the source speaker's F_0 contours depends on the degree and type of differences between the source and target contours. Given the pronounced intonation differences between the North American and (Glasgow) Scottish dialects [45, 11], it is perhaps no surprise that the linear model fared less well for speaker AWB. We need to take into account that the linear model as applied in this study did not accurately reflect the actual use in synthesis, in which the per-token mean and SD are — obviously — not given and where thus estimates need to be used. Thus, we do now know whether the latter actual-use method might have produced significantly worse results than the proposed method for speaker BDL, and not only, as was the case in the linear method employed in this study, for speaker AWB. Finally, our results may have implications for the role in speaker mimic of copying the mean and SD, or, in fact, of any approach based on copying statistical moments of theF_0 distribution and that does not take dynamic pattern into account.

7. REFERENCES

[1] E. E. Helander and J. Nurminen, "On the importance of pure prosody in the perception of speaker identity," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[2] E. Morley, E. Klabbers, J. P. van Santen, A. Kain, and S. H. Mohammadi, "Synthetic f0 can effectively convey speaker id in delexicalized speech." in *INTERSPEECH*, 2012.

[3] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for hmm-based speech synthesis using mllr," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 805–808.

[4] D. T. Chappell and J. H. Hansen, "Speaker-specific pitch contour modeling and modification," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2. IEEE, 1998, pp. 885–888.

[5] E. Helander and J. Nurminen, "A novel method for prosody prediction in voice conversion," in *Proceedings of ICASSP*, vol. 4, 2007, pp. 509–512.

[6] D. Lolive, N. Barbot, and O. Boeffard, "Pitch and duration transformation with non-parallel data," *Speech Prosody 2008*, pp. 111–114, 2008.

[7] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1145–1154, 2006.

[8] Z. Inanoglu and S. Young, "Emotion conversion using f0 segment selection." in *INTERSPEECH*, 2008, pp. 2122–2125.

[9] C. Veaux and X. Rodet, "Intonation conversion from neutral to expressive speech." in *INTERSPEECH*, 2011, pp. 2765–2768.

[10] J. Vaissière, "10 perception of intonation," *The handbook of speech perception*, p. 236, 2008.

[11] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008.

[12] P. Lieberman, "Intonation, perception, and language." *MIT Research Monograph*, 1967.

[13] J. P. Van Santen and B. Möbius, "A quantitative model of fo generation and alignment," in *Intonation*. Springer, 2000, pp. 269–288.

[14] G. K. Anumanchipalli, "Intra-lingual and cross-lingual prosody modelling," Ph.D. dissertation, Google Inc, 2013.

[15] J. P. van Santen, A. Kain, E. Klabbers, and T. Mishra, "Synthesis of prosody using multi-level unit sequences," *Speech Communication*, vol. 46, no. 3, pp. 365–375, 2005.

[16] J. P. van Santen, E. Klabbers, and T. Mishra, "Toward measurement of pitch alignment," *Italian Journal of Linguistics*, vol. 18, no. 1, p. 161, 2006.

[17] S. H. Mohammadi and A. Kain, "Semi-supervised training of a voice conversion mapping function using a joint-autoencoder," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[18] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 10, pp. 1506–1521, 2014.

[19] S. H. Mohammadi and A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 19–23.

[20] D. J. Patterson, "linguistic approach to pitch range modelling," Ph.D. dissertation, Edinburgh University, 2000.

[21] H. Fujisaki, "A model for synthesis of pitch contours of connected speech," *Annual Report, Engineering Research Institute, University of Tokyo*, vol. 28, pp. 53–60, 1969.

[22] M. S. Elyasi Langarani, E. Klabbers, and J. P. van Santen, "A novel pitch decomposition method for the generalized linear alignment model," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2584–2588.

[23] M. S. Elyasi Langarani, J. P. van Santen, S. H. Mohammadi, and A. Kain, "Data-driven foot-based intonation generator for text-to-speech synthesis." in *INTERSPEECH*, 2015.

[24] M. S. Elyasi Langarani and J. P. van Santen, "Modeling fundamental frequency dynamics in hypokinetic dysarthria," in *Spoken Language Technology (SLT), 2014 IEEE International Workshop on*. IEEE, 2014.

[25] G. Krishna Anumanchipalli, L. C. Oliveira, and A. W. Black, "Accent group modeling for improved prosody in statistical parameteric speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6890–6894.

[26] B. Gillett and S. King, "Transforming f0 contours," in *Proceedings of EUROSPEECH*, 2003, pp. 101–104.

[27] T. En-Najjary, O. Rosec, and T. Chonavel, "A new method for pitch prediction from spectral envelope and its application in voice conversion." in *INTERSPEECH*, 2003.

[28] M. Wang, M. Wen, K. Hirose, and N. Minematsu, "Emotional voice conversion for mandarin using tone nucleus model–small corpus and high efficiency," in *Proc. Speech Prosody*, 2012.

[29] Z. Inanoglu and S. Young, "A system for transforming the emotion in speech: combining data-driven conversion techniques for prosody and voice quality." in *INTERSPEECH*, 2007, pp. 490–493.

[30] A. Agarwal, A. Jain, N. Prakash, and S. Agrawal, "Word based emotion conversion in hindi language," in *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, vol. 9. IEEE, 2010, pp. 419–423.

[31] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken english," *Speech Communication*, vol. 51, no. 3, pp. 268–283, 2009.

[32] J. Ma and W. Liu, "Voice conversion based on joint pitch and spectral transformation with component group-gmm," in *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on*. IEEE, 2005, pp. 199–203.

[33] Z. Hanzlíček and J. Matoušek, "F0 transformation within the voice conversion framework," in *INTERSPEECH*, 2007.

[34] F.-L. Xie, Y. Qian, F. K. Soong, and H. Li, "Pitch transformation in neural network based voice conversion," in *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*. IEEE, 2014, pp. 197–200.

[35] A. Bayestehtashk and I. Shafran, "Parsimonious multivariate copula model for density estimation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 5750–5754.

[36] O. Türk and L. M. Arslan, "Voice conversion methods for vocal tract and pitch contour modification." in *INTERSPEECH*, 2003.

[37] C.-H. Wu, C.-C. Hsia, C.-H. Lee, and M.-C. Lin, "Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1394–1405, 2010.

[38] Z. Wu, T. Kinnunen, E. Chng, and H. Li, "Text-independent f0 transformation with non-parallel data for voice conversion." in *INTERSPEECH*, 2010, pp. 1732–1735.

[39] A. Black, P. Taylor, R. Caley, R. Clark, K. Richmond, S. King, V. Strom, and H. Zen, "The festival speech synthesis system, version 1.4. 2," *Unpublished document available via http://www. cstr. ed. ac. uk/projects/festival. html*, 2001.

[40] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 2, pp. 131–142, 1998.

[41] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1. IEEE, 1998, pp. 285–288.

[42] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[43] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk — a new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, January 2011.

[44] S. H. Mohammadi and A. Kain, "Transmutative voice conversion," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6920–6924.

[45] A. Cruttenden, *Rises in English*. In J. Windsor Lewis, editor, Studies in General and English Phonetics: Essays in Honour of Professor J. D. OĆonnor, Routledge, 1995.