INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Automated Emotion Morphing in Speech Based on Diffeomorphic Curve Registration and Highway Networks

Ravi Shankar[1], *Hsi-Wei Hsieh*[2], *Nicolas Charon*[2], *Archana Venkataraman*[1]

[1]Department of Electrical & Computer Engineering, Johns Hopkins University, Baltimore MD, USA
[2]Department of Applied Mathematics & Statistics, Johns Hopkins University, Baltimore MD, USA

rshanka3@jhu.edu, {hhsieh,charon}@cis.jhu.edu, archana.venkataraman@jhu.edu

Abstract

We present a novel approach for emotion conversion that bridges the domains of speech analysis and computer vision. Our strategy is to warp the pitch contour of a source emotional utterance using diffeomorphic curve registration. The associated dynamical process pushes the original source contour towards that of a target emotional utterance. Mathematically, this warping process is completely specified by a set of *initial momenta*. Therefore, we use parallel data to train a highway neural network (HNet) to predict these initial momenta directly from the signal characteristics. The input features to the HNet include contextual pitch and spectral information. Once trained, the HNet is used to obtain the initial momenta for new utterances. From here, the diffeomorphic process takes over and warps the pitch contour accordingly. We validate our framework on the VESUS repository collected at Johns Hopkins University, which contains parallel emotional utterances from 10 actors. The proposed warping is more accurate that three state-of-the-art baselines for emotion conversion. We also evaluate the quality of our emotion manipulations via crowd sourcing.

Index Terms: Emotional speech morphing, 2D curve registration, momentum estimation, highway neural network

1. Introduction

Human speech contains a vast amount of information beyond the semantic content. For example, the manner of speaking implicitly reflects our emotional state and intent [1, 2]. While humans are adept at generating and parsing these emotional cues, the same cannot be said for automated platforms. One reason is that emotions are highly complex with overlapping signal attributes, thus making them difficult to disentangle. Another reason is the lack of freely available emotional speech data to train end-to-end systems. As a result, most emotion recognition models cannot generalize beyond individual datasets, and expressive synthesis remains an open problem [3]. This paper circumvents the challenges of prior work by focusing on the problem of *emotion conversion*. Namely, given a neutral speech utterance, we learn a model to transform the emotional content without altering the semantic or speaker information.

At a high level, emotional speech is controlled by three prosodic attributes: pitch also known as fundamental frequency (F0), signal intensity, and speaking rhythm [1]. Out of these, the pitch contour controls intonation, which plays a crucial role in emotional expression. For example, anger is often characterized by sharp increases in pitch, while sadness is linked to gradual pitch reductions. Several previous works have explored the problem of emotion conversion via prosodic manipulation. A explicit modeling of the pitch contour was proposed by [4]. The authors compared the accuracy of linear regression, a Gaussian mixture model (GMM) and nonlinear regression trees in performing the required manipulations. Another method proposed by [5] independently modified prosodic and spectral features. Similarly, a GMM with global variance constraint was proposed for voice conversion in [6] and later adopted for emotion conversion by [7]. Sparse coding and dictionary learning based strategies have also been used for emotion conversion. In particular, the work of [8] developed a non-negative matrix factorization (NMF) model to learn parallel spectral dictionaries for the source and target emotions. A new utterance is first encoded using the source dictionary and reconstructed using the target. With the advent of deep learning, the work of [9] proposed a bi-directional long-short memory network (Bi-LSTM) model for emotion conversion, which acts on both the prosodic features and a parameterized spectrum. The prosodic features are also affected by both short-term (phoneme level) and long-term (syllables or words level) acoustic events of an utterance.

This paper develops a completely novel framework for emotion conversion based on the principles of deformable image registration. Image registration is a widely studied problem in computer vision, where the goal is to align two images by manipulating the underlying coordinate systems [10, 11]. Within this class of transformations, diffeomorphic algorithms are based on a smooth and invertible displacement field using an exponential mapping [12]. In our case, the "images" will correspond to 2-D pitch contours of the same utterance in the source and target emotions. We will learn a simple vertical transformation that locally changes the pitch without affecting the speaker identity or speaking rate. Mathematically, this transformation is parameterized by an initial displacement field, also known as the *momenta*. Our strategy is to predict the initial momenta using a highway neural networks (HNet) architecture. The HNet input features consist of the raw spectrogram averaged within the standard Mel-frequency bands, along with the F0 values in a 400 ms context window. We train and evaluate our model on the VESUS emotional speech database collected at Johns Hopkins [13]. We compare the performance of our momenta prediction algorithm with three state-of-the-art baselines.

2. Momentum-Based Emotion Conversion

VESUS contains parallel emotional utterances, which allows us to draw frame-wise correspondences. We use the STRAIGHT vocoder [14] to extract pitch contours from the utterances. During training, we align the source and target emotional utterances using dynamic time warping [15]. From here, we use the formulation in the next section to estimate the frame-wise momenta for each utterance pair. We then train an HNet to predict these momenta based on the pitch and spectral information in the original utterance. During testing, we estimate the frame-wise pitch momentum using our trained HNet and apply the diffeo-

http://dx.doi.org/10.21437/Interspeech.2019-2386

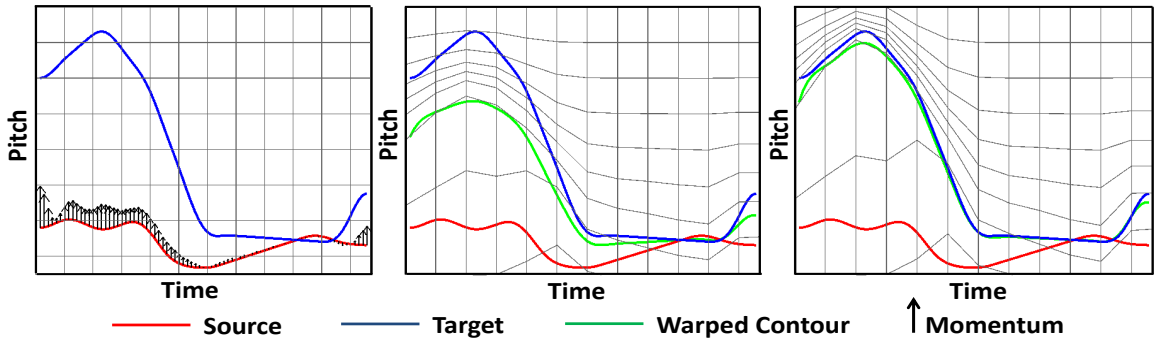

Figure 1: *Illustration of 2-D diffeomorphic registration for emotion conversion.* ***Left:*** *source (neutral) and target (emotional) pitch contours from parallel utterances.* ***Middle:*** *intermediate output as source moves towards target.* ***Right:*** *final curve alignment.*

morphic transformation to obtain the new pitch contour. We resynthesize the modified utterance again using STRAIGHT.

2.1. Diffeomorphic Registration for 2-D Curves

Our goal in this work is to learn a *transformation* on pitch contours that alters the perceived emotional content of the reconstructed utterance. We adopt the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework [16, 17], which provides global convergence and optimality guarantees. At a high level, LDDMM is based on an underlying vector field that acts on the source contour. This vector field is parameterized by an exponential map, which provides a smooth transition. For simplicity, we assume that the signals have been aligned using dynamic time warping (DTW). In this case, the vector field acts only in the vertical direction to locally change the pitch values. Fig. 1 illustrates this warping process on two pitch contours.

Mathematically, let $\mathbf{p_t}$ and $\hat{\mathbf{p}}_\mathbf{t}$ be the source and target pitch contours, respectively. The time index $\mathbf{t}$ corresponds to the discrete sampling of the contours from $\mathbf{t} = 0, \ldots, T$. Our approach is related to the landmark LDDMM setting of [18, 19, 20] with a vertical constraint on the vector field. In particular, let $\mathbf{v_t}(\mathbf{x}; s)$ be a non-stationary and finite norm vector field across time $\mathbf{t}$ and pitch values $\mathbf{x}$. These vector fields generate the dynamical deformations with respect to the second evolution argument s. Namely, for a fixed point in time $\mathbf{t}$, we can consider the continuous flow $\mathbf{x} \mapsto \varphi^{\mathbf{v}}_{\mathbf{t}}(\mathbf{x}; s)$ of the vector field for $s \in [0, 1]$ defined by $\varphi^{\mathbf{v}}_{\mathbf{t}}(\mathbf{x}; 0) = \mathbf{p_t}$ and the ordinary differential equation (ODE) $\partial_s \varphi^{\mathbf{v}}_{\mathbf{t}}(\mathbf{x}; s) = \mathbf{v_t}(\varphi^{\mathbf{v}}_{\mathbf{t}}(\mathbf{x}; s); s)$. Here, the initial condition specifies that we begin the evolution process from the source pitch contour. The ODE specifies that the displacement at every new pitch value is given by the vector field $\mathbf{v_t}(\mathbf{x}; s)$. The evolution process terminates at $s = 1$.

We now formulate the registration problem between the source pitch contour $\mathbf{p_t}$ and the target pitch contour $\hat{\mathbf{p}}_\mathbf{t}$ through the following optimal control problem:

$$\min_{\mathbf{v} \in \mathcal{V}} \frac{1}{2} \int_0^1 \|\mathbf{v_t}(\cdot; s)\|_V^2 \, ds + \lambda \sum_{t=1}^{T} (\varphi^{\mathbf{v}}_{\mathbf{t}}(\mathbf{p_t}; 1) - \hat{\mathbf{p}}_\mathbf{t})^2 \quad (1)$$

The first term of Eq. (1) is a smoothness constraint on the underlying vector field. The Hilbert norm $||\cdot||_V$ is implicitly defined through a 2-D exponential kernel that operates across time and pitch. The second term of Eq. (1) is the data matching term, which enforces that the warped source contour should be close to the target contour. Notice that the parameter λ controls the trade-off between smoothness and registration fidelity.

The Pontryagin maximum principle of optimal control [20] allows us to derive necessary conditions for the solution to Eq. (1). In this case, the theory shows that there exist variables $\mathbf{m}^s_\mathbf{t}$ for $s \in [0, 1]$ that we call *momenta*. These momenta behave like hidden state variables in the continuous-time Kalman filter framework. The "observed" variables in this analogy are the pitch values of the warped contour. The Hamiltonian dynamics associated with the state/observer model allow us to reformulate Eq. (1) as a minimization over initial momenta $\mathbf{m}^0_\mathbf{t}$.

Formally, let $\mathbf{z}_t(s) = [\mathbf{t} \;\; \varphi^{\mathbf{v}}_{\mathbf{t}}(\mathbf{p_t}; s)]^{\mathrm{T}}$ be a two-dimensional vector of the time and deformed pitch value, and let $\gamma_{ij}(s)$ be the kernel evaluated at the pair of vectors $\mathbf{z}_i(s)$ and $\mathbf{z}_j(s)$. The quadratic objective for the collection of initial momenta can be written as follows:

$$\mathcal{J}(\mathbf{m}^0) = \frac{1}{2} \sum_{i,j=1}^{T} \gamma_{ij}(0) \mathbf{m}^0_\mathbf{i} \mathbf{m}^0_\mathbf{j} + \lambda \sum_{t=1}^{T} (\varphi^{\mathbf{v}}_{\mathbf{t}}(\mathbf{p_t}; 1) - \hat{\mathbf{p}}_\mathbf{t})^2 \quad (2)$$

subject to Hamiltonian equations. A standard approach to solve such a problem numerically is given by *shooting algorithms* [21]. We essentially apply a quasi-Newton descent method on $\mathcal{J}$, where the gradient w.r.t $\mathbf{m}^0$ of the second term in Eq. (2) is computed via the adjoint Hamiltonian equations.

Our strategy is to use Eq. (2) to solve directly for the initial momenta in the training dataset, where we have access to parallel emotional utterances. We will then train a neural network to predict these momenta directly from the signal characteristics. This neural network will be applied to the testing utterances to predict the (unknown) initial momenta. The contour registration process is completely specified once we have these values.

2.2. Input Features for Momentum Prediction

As described above, our model predicts the initial displacement (i.e., momenta) to transform a source utterance to the target emotion. We use two classes of features to predict the frame-wise momentum: a compressed form of the raw spectrum and the original pitch contour with a 200 ms context on both sides of the frame. Our rationale for using a long contextual window for pitch is to account for both local and global properties. Since pitch is affected by both segmental (phonetic level) and supra-segmental (syllable or word level) characteristics, a context of 360 ms ensures that the pitch information is provided over on average two syllables. All input features are extracted using a frame period of 5 ms and a 5 ms window stride.

To reduce the input dimensionality, we compress the raw spectral envelope using the normalized Mel frequency. Specifi-

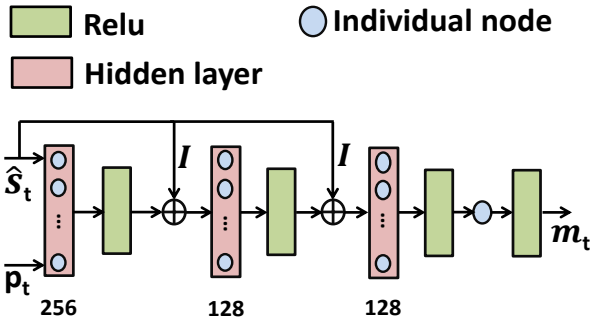

Figure 2: *HNet architecture for initial momentum prediction.*

cally, we first compute a 1,024 point FFT for each time frame, which results in a 513 dimensional magnitude spectrum $F_t \in \mathbb{R}^{513\times 1}$ (frequency range 0 to π). We use the normalized Mel filterbank matrix to obtain a 128-dimensional input representation $\hat{S}_t \in \mathbb{R}^{128\times 1}$. The filterbank matrix preserves the shape of the spectrum while preserving the acoustic information present in the frame. Our compression scheme is highly effective in accelerating the training times for our deep neural networks. Empirically, we find that further compression beyond 128 dimensions leads to undesirable distortions in the spectral envelope.

2.3. Highway Neural Network Architecture

We employ an artificial neural net with skip connections between the input and hidden layers. This architecture is known as a highway network (HNet). Our model contains one input layer, three hidden layers, and one output layer, as illustrated in Fig. 2. The input spectral features $\hat{\mathbf{s}}_\mathbf{t}$ are normalized to mean 0 and unit variance while the pitch contours $\mathbf{p_t}$ are fed in without any normalization. The output of neural network, i.e., the initial momentum $\mathbf{m_t}$, is given by the following expression:

$$\begin{aligned}\mathbf{m_t} = \phi[W_{34} \times \phi[W_{23} \times (\phi[W_{12} \\ \times (\phi[W_{01} \times \{\hat{\mathbf{s}}_\mathbf{t}, \mathbf{p_t}\} + b_1] \\ \oplus \mathbf{I}\hat{\mathbf{s}}_\mathbf{t}) + b_2] \oplus \mathbf{I}\hat{\mathbf{s}}_\mathbf{t}) + b_3] + b_4] \quad (3)\end{aligned}$$

The variables W_{ij} in Eq. (3) denote the weights going from layer i to layer j, and ϕ is the ReLU non-linearity [22] applied at each hidden layer and the output. The variable b_i is the bias related to the layer i. The term $\mathbf{I}\hat{\mathbf{s}}_\mathbf{t}$ denotes the skip connections concatenated to the second and third hidden layer output, respectively. The variable $\mathbf{I}$ is the identity matrix showing there is no transformation of the features being carried out in skip connections. Variable $\mathbf{m_t}$ is the momentum predicted for the input source frame t. We use a dropout [23] rate of 0.3 and batch normalization [24] after every hidden layer and before the skip connections with identity. We use the Adam optimizer [25] with a fixed learning rate of 0.01 and mini-batch sizes of 500.

2.4. Reconstruction

The predicted momenta are used to transform the entire source pitch contour. The aperiodicity and spectrogram components are copied directly from the source speech. We reconstruct the modified utterance using STRAIGHT by replacing the source pitch contour with the transformed version.

3. Experimental Setup

We performed both an objective and subjective evaluation of our momentum prediction framework. The results are compared to three state-of-the-art emotion conversion baseline algorithms.

3.1. Emotional Speech Dataset and Evaluation

Our training and evaluation relies on the VESUS emotional dataset collected at Johns Hopkins University [13]. VESUS contains parallel emotional utterances spoken by a mix of amateur and professional actors. The database has 2500 utterances for each of five emotional classes: happiness, anger, sadness, fear and neutral. The dataset also contains perception ratings for each utterance provided by 10 raters on Mechanical Turk.

In this work, we consider three emotion conversion models: neutral to angry, neutral to sad, and neutral to happy. These conversions span both high- and low-arousal emotions to test the limits of our diffeomorphic registration approach. We also sub-select the VESUS utterances based on $\geq 50\%$ agreement between raters. The total numbers in our experiment are:

- **Neutral to Angry:** 1534 utterances for training, 72 for validation, and 100 for testing.
- **Neutral to Happy:** 790 utterances for training, 43 for validation, and 43 for testing.
- **Neutral to Sad:** 1449 utterances for training, 63 for validation, and 70 for testing.

Our objective evaluation includes the mean absolute error and the Pearsons correlation coefficient measure between the predicted pitch values and their corresponding ground truth counterparts. For subjective evaluation, we ask human raters on AMT (Amazon Mechanical Turk) to score each of the converted test sample for perceived emotion. The survey plays two audio files for the raters to listen. One of them is the baseline neutral speech and the other one is the speech converted into one of the target emotions. The order of neutral and emotional speech is randomized in each trial to weed out any non-diligent raters. After they are done listening, we ask them to independently classify the emotion in both audio files. A bias correction using source (neutral) speech is important in our evalution because emotion perception is highly dependent on knowledge about the speaker articulation or manner of speaking.

3.2. Baseline methods

We compare the momentum prediction model against three state-of-the-art baseline methods for emotion conversion. The first baseline fits a Gaussian mixture model (GMM) to the joint distribution of the source and target STRAIGHT cepstrum features and fundamental frequency [7]. We use the global variance constraint proposed by [6] to improve the GMM accuracy.

The second baseline relies on the dictionary learning and sparse Non-Negative Matrix Factorization (NMF) method developed in [8]. Here, two parallel dictionaries of STRAIGHT spectrum are constructed from the training dataset by using an active Newton set based method. NMF estimates the sparse coding of the input spectral features over the source dictionary. This sparse coding is then used to construct the converted spectrum and fundamental frequency using the target dictionary.

The third baseline is the Bi-LSTM model developed in [26]. As outlined in the original publication, we pre-train the Bi-LSTM on the CMU-ARCTIC voice conversion corpus [27] and fine-tune it for emotion conversion on the VESUS database. This method simultaneously converts both spectral and prosody

Table 1: *MAE and Pearson's Correlation measures for pitch across target emotions using multi-speaker model.*

Algorithm	MAE(F0)	Corr(F0)
Neutral-to-Angry		
GMM	44.3	0.54
NMF	94.2	0.22
Bi-LSTM	57.4	0.34
Proposed	**40.5**	**0.61**
Neutral-to-Happy		
GMM	53.8	0.51
NMF	106.7	0.25
Bi-LSTM	67.6	0.48
Proposed	**49.8**	**0.54**
Neutral-to-Sad		
GMM	29.1	**0.8**
NMF	65.3	0.4
Bi-LSTM	29.6	0.78
Proposed	**27.7**	0.74

(pitch, energy) features between source and target emotion. The prosody features are parameterized by a continuous wavelet transform [28]. The intention behind such parameterization is to consider both short-term and long-term pitch and energy trajectories by using ten different wavelet scales.

4. Experimental Results

Table 1 summarizes the objective results obtained for baseline and proposed methods. Our algorithm is uniformly better at approximating the target pitch contour in absolute error sense. The results demonstrate that our parameterization of pitch deformation by intial momentum does work effectively.

The GMM based prosody and spectrum conversion comes a close second, beating both NMF and Bi-LSTM based models. The reason for this can be attributed to the simplicity of GMM which allows it to learn the parameters i.e., mean and covariances in high dimensional space. However, the speech reconstructed by GMM is poor because of the averaging effect that mixture models have. It fails to conditionally sample from the tails of joint distribution and hence the predicted pitch wiggles about the mean of the training data. NMF does a poor job in prediction of prosody because of the lack of any global constraint while estimating sparse coding. The cepstral features are not a unique representation of an acoustic unit and there exist a many-to-one mapping. This further results in discontinuities in the converted spectrum going from one frame to the next. In the end, the reconstructed speech is very distorted and sometimes completely unitelligible. Bi-LSTM does worse compared to our method of pitch approximation because of its over parameterization. The multi-scale wavelet transform used for encoding the prosodic features leads to a very rough estimate of the predicted pitch and energy contour. Furthermore, the underlying assumption about the existence of local minima for emotion conversion being close to the voice conversion optima is not always true.

In contrast, our proposed model predicts only one value which is the initial momentum parameter. Besides, we design

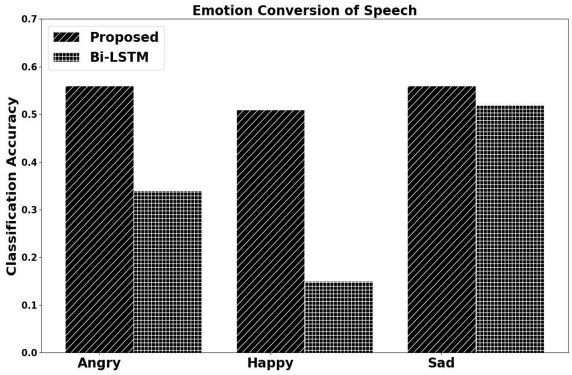

Figure 3: *Comparison of emotion classification accuracy.*

our HNet to appropriately learn this regression function by minimizing the l_1 penalty which, unlike l_2 loss allows the model to evenly focus on the less extreme parts of the target distribution.

Our subjective evaluations are based on five crowd-sourced ratings for each converted speech via AMT. A majority voting decides the final emotion label of the converted utterances. We found the reconstructed speech from the GMM and NMF models are be highly distorted and unintelligible. Therefore, we only obtain crowd-sourced ratings for our HNet and the Bi-LSTM model. To get a uniform comparison between the proposed method and Bi-LSTM based conversion, we crowdsource the ratings for exact same utterances spoken by same speakers. Fig.3 shows the emotion classification accuracy on the testing utterances. Compared to the baseline model, our proposed model has higher classification accuracy across all three emotions. Further, the classification for neutral-to-angry is the best followed by neutral-to-happy and then neutral-to-sad. Comparatively, the high arousal emotions like angry or happy are easier to discern than low arousal emotions like sad. This effect is evident in the Fig.3 as the difference in classification accuracy is lowest for neutral-to-sad conversion. Our method, unlike the Bi-LSTM model, only modifies the pitch and still does remarkably better on the listening tasks. This proves that the proposed method is very robust for carrying out emotion morphing. Another point to be noted is that, since we only modify the pitch and not the spectral envelope, the speaker information is retained and the converted speech is distortionless.

5. Conclusion

We proposed a method for emotion conversion based on estimating a curve warping function for pitch contours. The warping was based on a diffeomorphic registration technique that generates a sequence of smooth and invertible time-varying vector fields in an iterative fashion. We trained a highway network to predict the deformation parameter, also called as the initial momentum, for every point on a given pitch contour. The warped curve was used to reconstruct speech for three target emotions. Our experiments showed that the speech generated by modified pitch contours were perceived more emotional than speech generated by the baseline algorithm. Furthermore, our proposed model retained the speaker characteristics and the quality of speech by not changing the spectral envelope of the source audio. As a future direction, we plan to modify both pitch and speaking rate (duration) to exercise a control over the strength of target emotion in converted speech.

Acknowledgements: We thank Jacob Sager for his help with the crowd sourcing experiments.

6. References

[1] T. Johnstone and K. Scherer, "Vocal communication of emotion," *Handbook of Emotions,*, 01 2000.

[2] D. Schacter, D. T. Gilbert, and D. M. Wegner, *Psychology (2nd Edition)*. New York: Worth, 2011.

[3] M. Avanzi, G. Christodoulides, D. Lolive, E. Delais-Roussarie, and N. Barbot, "Towards the adaptation of prosodic models for expressive text-to-speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 09 2014.

[4] Y. Kang, J. Tao, and B. Xu, "Applying pitch target model to convert f0 contour for expressive mandarin speech synthesis," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 1, 06 2006, pp. I – I.

[5] Z. Inanoglu and S. Young, "A system for transforming the emotion in speech: Combining data-driven conversion techniques for prosody and voice quality," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 1, 01 2007, pp. 490–493.

[6] 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.

[7] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Gmm-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing*, vol. 2, pp. 134–138, 12 2012.

[8] R. Aihara, T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 05 2014, pp. 7894–7898.

[9] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *Transaction on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[10] W. Li and H. Leung, "A maximum likelihood approach for image registration using control point and intensity," *IEEE Transactions on Image Processing*, vol. 13, no. 8, pp. 1115–1127, Aug 2004.

[11] L. G. Brown, "A survey of image registration techniques," *ACM Computation Survey*, vol. 24, no. 4, pp. 325–376, Dec. 1992.

[12] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Non-parametric diffeomorphic image registration with the demons algorithm," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007*, N. Ayache, S. Ourselin, and A. Maeder, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 319–326.

[13] J. Sager, R. Shankar, J. Reinhold, and A. Venkatarman, "Vesus: A crowd-annotated database to study emotion production and perception in spoken english," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019.

[14] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveign, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 04 1999.

[15] *Dynamic Time Warping (DTW)*. Dordrecht: Springer Netherlands, 2008, pp. 570–570.

[16] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *International journal of computer vision*, vol. 61, no. 139-157, 2005.

[17] L. Younes, *Shapes and diffeomorphisms*. Springer, 2010.

[18] S. C. Joshi and M. I. Miller, "Landmark matching via large deformation diffeomorphisms," *IEEE transactions on image processing*, vol. 9, no. 8, pp. 1357–1370, 2000.

[19] M. Miller, A. Trouvé, and L. Younes, "Hamiltonian Systems and Optimal Control in Computational Anatomy: 100 Years Since D'Arcy Thompson." *Annual Review Biomedical Engineering*, vol. 7, no. 17, pp. 447–509, 2015.

[20] S. Arguillère, E. Trélat, A. Trouvé, and L. Younes, "Registration of Multiple Shapes using Constrained Optimal Control," *SIAM Journal on Imaging Sciences*, vol. 9, no. 1, pp. 344–385, 2016.

[21] F.-X. Vialard, L. Risser, D. Rueckert, and C. J. Cotter, "Diffeomorphic 3d image registration via geodesic shooting using an efficient adjoint calculation," *International Journal of Computer Vision*, vol. 97, no. 2, pp. 229–241, 2012.

[22] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10. USA: Omnipress, 2010, pp. 807–814.

[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[26] H. Ming, D.-Y. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional lstm modeling of timbre and prosody for emotional voice conversion," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 09 2016, pp. 2453–2457.

[27] J. Kominek and A. W Black, "The cmu arctic speech databases," *ISCA Speech Synthesis Workshop-2004*, 01 2004.

[28] M. Sam Ribeiro and R. A. J. Clark, "A multi-level representation of f0 using the continuous wavelet transform and the discrete cosine transform," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 04 2015, pp. 4909–4913.