NEUTRAL TO LOMBARD SPEECH CONVERSION WITH DEEP LEARNING

Enguerrand Gentet[1,2]*, Bertrand David*[1]*, Sébastien Denjean*[2]*, Gaël Richard*[1]*, Vincent Roussarie*[2]

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2]Groupe PSA, Chemin de Gisy, 78943 Vélizy-Villacoublay, France

ABSTRACT

In this paper, we propose several approaches for neutral to Lombard speech conversion. We study in particular the influence of different recurrent neural network architectures where their main hyper-parameters are carefully selected using a bandit-based approach. We also apply the Continuous Wavelet Transform (CWT) as a multi-resolution analysis framework to better model temporal dependencies of the different features selected. The speech conversion results obtained are validated by means of objective evaluations which highlight in particular the interest of the wavelet transform for the learning process.

Index Terms— speaking style conversion, lombard effect, deep learning, recurrent neural networks, wavelets

1. INTRODUCTION

Speaking Style Conversion (SSC) aims at modifying the style of a given speech signal while keeping the speaker acoustic characteristics. This is of clear interest e.g. for the improvement of current speech synthesis systems in allowing a wide range of personalized and specific speaking styles. Some examples of SSC include the conversion of neutral to emotional speaking styles (happy, sad, fear, angry) [1–6] or to vocal effort speaking styles [7–11].

One particular vocal effort based speaking style called Lombard effect [12] refers to the speech changes involuntary induced by a speaker while he is communicating in noise. This is particularly interesting since a Neutral-to-Lombard SSC system has the potential to improve the intelligibility of the original synthetic (or natural) speech in noisy environments. For example, adapting the neutral speech of car on-board applications or announcement systems in large train stations may increase the overall intelligibility of the message and induce an improved end-user hearing experience. One of the main interests in such a transformation is to adapt the rendered speech in the listening environment (car, train station, ...) taking into account the surrounding noise while the speech (natural or synthetic) has been produced in a different or silent environment.

SSC transformations can be either direct or parametric. The direct approach aims at filtering the signal as in [7]. The parametric approach is based on a model such as a vocoder to extract speech parameters that will then be transformed [1–6, 8–10, 13]. This latter approach allows a more comprehensive processing while maintaining a good quality and naturalness of the converted speech. In this paper, we follow the same strategy and use a parametric approach based on the well known STRAIGHT vocoder [14], widely used in SSC [1–3, 5, 6, 13]. This has been shown to be a good compromise between quality of speech and perceptive changes for the neutral to Lombard SSC task among other vocoders [9].

In early works the functions used to transform the speech parameters are empirical, based on rules made out of observations of the speaking style [4, 13]. The results are interesting but important temporal and inter-features dependencies are not well taken into account. Hence, most recent SSC approaches are based on statistical learning models that learn the speech features transformation functions exploiting more or less large speech dataset. Previous methods exploit Gaussian Mixture Models (GMM) [1–3, 5, 8], Deep Neural Network (DNN) architectures [9] and more recently RNN architectures [6]. Most approaches are based on parallel data learning techniques, with some exceptions such as for example with Cycle consistent Generative Adversarial Networks (CycleGANs) [10]. To the best of our knowledge, RNN architectures have not been used for neutral to Lombard SSC. In this paper we study the influence of different RNN architectures and its main hyper-parameters on the learning task in SSC by focusing on the neutral to Lombard case. The hyper-parameters are carefully selected using a bandit-based approach called Hyperband [15] to select nearly optimal settings for each proposed model.

Another contribution of this work is to consider a multi-resolution analysis framework to better model temporal dependencies of the different features selected. In fact, the use of the Continuous Wavelet Transform (CWT) has been suggested to describe speech parameters at several time-scales [16] and already used with success in emotional SSC [5, 6]. In this paper, we more specifically study the effect of CWT on the fundamental frequency and the energy contour features applied to the neutral to Lombard SSC task.

Our results show that the use of CWT greatly improves the learning process for the fundamental frequency feature but it is less noticeable for the energy contour feature. Some RNN variants also manage to slightly enhanced the learning task for every speech feature.

The paper is organized as follows. In section 2, we detail the SSC system framework and speech features considered in our study. The system performance in section 3. We finally suggest some conclusions and perspectives of this work.

2. SPEECH STYLE CONVERSION (SSC) SYSTEM

2.1. Framework

A block diagram describing the information flow of our system can be seen on figure 1. In the training phase the vocoder first analyses the speech parameters for each utterance in the source and target styles. The parameters to map are pre-processed into features then aligned with Dynamic Time Warping (DTW). Finally, the learning model is trained using the resulting parallel features. Once the model is trained, the framework can be used to convert new speech signals from the source style to the target style. In this generative phase, only the features of the source style speech are computed and fed to the trained model to estimate the converted features. A reconstruction

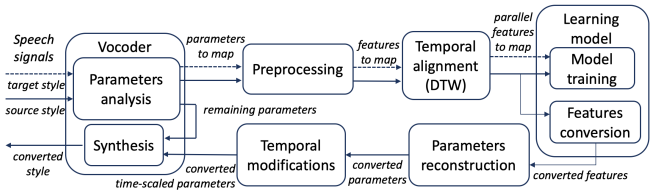

Fig. 1. Block diagram of our SSC system.

step then allows to get the converted speech parameters that are used by the vocoder to synthesize the converted speech. Finally, temporal modifications may be applied before the synthesis.

2.2. Speech parameters and features

The learning features are computed from the speech parameters so they heavily depend on the vocoder used. In this study we are using the STRAIGHT source-filter vocoder [14] to extract the speech parameters including the power spectrum, aperiodicity spectrum and fundamental frequency. All the parameters are computed frame-wise with a hop duration of 5 ms.

Since our interest is to alter the timbre and prosody of the speech signal, we rely on features classically used in SSC systems. These features are the logarithmic values of the fundamental frequency, the logarithmic values of the energy contour, and the Mel-Frequency Cepstral Coefficients (MFCCs). The fundamental frequency values of the input signal segmented in L frames, noted $\mathbf{f0} \in \mathbb{R}^{1\times L}$, are directly obtained with STRAIGHT. The energy contour values, noted $\mathbf{e} \in \mathbb{R}_+^{1\times L}$, are computed from the STRAIGHT power spectrum, noted $|\mathbf{S}|^2 \in \mathbb{R}_+^{N\times L}$, as follows:

$$\mathbf{e}_i = \sqrt{\sum_{n=0}^{N-1} |\mathbf{S}_{i,n}|^2}\,. \tag{1}$$

In order to manipulate perceptually relevant magnitudes, the fundamental trajectory is transformed to the logarithmic semi-tone scale $\mathbf{f0}^{st} = 39.87 \log_{10}(\mathbf{f0}/50)$ and the energy contour on the decibel scale $\mathbf{e}^{dB} = 20 \log_{10} \mathbf{e}$. The MFCCs, noted $\mathbf{mc} \in \mathbb{R}^{M\times L}$, are conventionally computed applying a M-dimensional mel-spaced filterbank on the power spectrum then taking the Discrete Cosine Transform (DCT) from the log of the resulting mel-power spectrum :

$$|\mathbf{S}|^2 \in \mathbb{R}_+^{N\times L} \xrightarrow{\text{mel scale}} |\tilde{\mathbf{S}}|^2 \in \mathbb{R}_+^{M\times L} \xrightarrow{\log + \text{DCT}} \mathbf{mc} \in \mathbb{R}^{M\times L}\,. \tag{2}$$

We choose $M = 25$ but we do not retain the first coefficient since it is directly related to the energy of the analysed frame. For the generative phase, we simply use an inverse DCT to convert the mel-cepstrum back to the power spectrum.

In this work a Continuous Wavelet Transform (CWT) is used to describe the fundamental frequency and energy contour at several time-scales with an objective of improving the learning process for these features. The continuous wavelet transform of a discrete sequence $\mathbf{x} \in \mathbb{R}_+^{1\times L}$ at a scale $s \in \mathbb{R}_+^*$ is computed as follows :

$$\mathbf{W}_i(s) = \sum_{i'=0}^{L-1} \mathbf{x}_{i'} \psi^* \left[\frac{i'-i}{s}\right] \tag{3}$$

where $\psi(t)$ is a function called the mother wavelet. First, a linear interpolation is used on $\mathbf{f0}^{st}$ to fill the unvoiced segments. Then $\mathbf{f0}^{st}$ and $\mathbf{e}^{dB}$ are standardized to zero mean and unit variance as required by wavelet analysis. Finally a ten octave scale CWT is applied on both features using the classic Ricker mother wavelet (sometimes called Mexican hat) of length $s_0 = 2$ i.e. a temporal duration of 10 ms. The features noted $\mathbf{x}^{cwt} \in \mathbb{R}^{10\times L}$ are represented by the resulting components :

$$\mathbf{x}_{n,i}^{cwt} = \mathbf{W}_i(s_n) = \mathbf{W}_i(2^n s_0)\,. \tag{4}$$

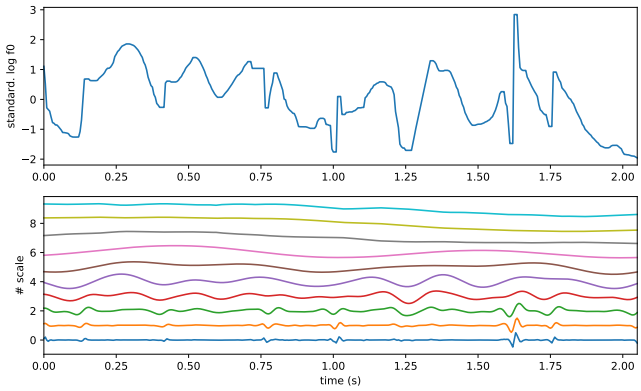

Fig. 2. Example of a standardized $\mathbf{f0}^{st}$ and its CWT coefficients.

An example of standardized $\mathbf{f0}^{st}$ and its CWT components can be seen on figure 2. For the generative phase, the reconstruction is achieved using the following equation given in [17] :

$$\hat{\mathbf{x}}_i = \frac{1}{C_\delta \psi_0(0)} \sum_{n=0}^{9} \frac{\mathbf{x}_{n,i}^{cwt}}{\sqrt{s_n}}\,, \tag{5}$$

where $\psi_0(0) = 0.8673$ is the normalized wavelet basis function evaluated at time zero and $C_\delta = 3.541$ is the given reconstruction factor for the Ricker wavelet.

Finally, from any feature $\mathbf{x} \in \mathbb{R}_+^{1\times L}$ one can compute its delta (resp. delta-delta), features noted $\mathbf{x}^\delta$ (resp. $\mathbf{x}^{\delta^2}$), as follows :

$$\mathbf{x}_i^\delta = \text{delta}(\mathbf{x}_i) = \frac{\sum_{d=1}^{D} d(\mathbf{x}_{i+d} - \mathbf{x}_{i-d})}{2\sum_{d=1}^{D} d^2}\,, \tag{6}$$

$$\mathbf{x}_i^{\delta^2} = \text{delta}(\mathbf{x}_i^\delta)\,, \tag{7}$$

where typically D = 2. For the generative phase, to take the dynamics of a feature into account, the Maximum Likelihood Parameter Generation (MLPG) algorithm [18] is used to estimate the feature trajectory :

$$\hat{\mathbf{x}} = \text{MLPG}(\mathbf{x}, \mathbf{x}^\delta, \mathbf{x}^{\delta^2}) \tag{8}$$

2.3. Learning models

The learning model chosen has a major influence on the system performance. In this work we study the influence of several architectures in SSC applied to the neutral to Lombard speech conversion task from the usual Gaussian Mixture Models (GMM) to more recent Deep Neural Network (DNN) architectures such as Feed-Forward Neural Network (FFNN) and Recurrent Neural Network (RNN). More specifically, we study the influence of three popular RNN

		FFNN		RNNs	
hp	distributions	low	high	low	high
N_{hl}	int uniform	2	6	2	4
N_{hu}	int log-uniform	64	512	32	512
dr	uniform	0.	0.4	0.	0.4
N_{batch}	int log-uniform	128	1024	1	32
lr	log-uniform	10^{-5}	10^{-3}	10^{-5}	10^{-3}
wd	log-uniform	10^{-9}	10^{-5}	10^{-9}	10^{-5}

Table 1. Probability distributions of the hyper-parameters. The distributions may be discrete (int), uniform or log-uniform.

variants, namely Fully Recurrent (FR), Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM). All these variants are Uni-Directionnal (UD) as they predict each element of the sequence based on the previous ones but they may be used as Bi-Directional (BD) RNN to predict each element also based on the next ones.

The choice of the hyper-parameters is critical on a model performance. For the GMM we use a variant called Variational Bayesian GMM (VBGMM) with a high number of mixtures (200) because VBGMM has a natural tendency to set some mixture weights values close to zero and then automatically choose a suitable number of effective components. However for the DNN models, hyper-parameters are carefully selected using a bandit-based approach called Hyperband [15]. It is a recent adaptive random search that managed to provide a speedup over other hyper-parameters search algorithms on a variety of deep-learning problems.

This algorithm is based on adaptive resource allocation and early-stopping : several brackets of randomly selected configurations are run with a limited amount of training resource allocated before some configurations are discarded; a larger number of configurations in a bracket corresponds to a smaller allocated resource and hence more aggressive early-stopping. For every DNN model we are running Hyperband with the suggested input values i.e $R = 81$ and $\eta = 3$ where R is the maximum amount of resource allocated to a configuration and $1/\eta$ is the proportion of configurations discarded in each round.

All the DNN models are using N_{hl} hidden layers with N_{hu} hidden neurons each, with a dropout rate of dr. The batch size is N_{batch} and the optimization algorithm used is Adam with a starting learning rate of lr and a weight decay of wd. The FFNN hidden neurons are Rectified Linear units (ReLu). Preliminary tests allowed us to choose ranged probability distributions, displayed in table 1, that sample the hyper-parameter search space for every model.

3. SYSTEM PERFORMANCE

3.1. Experimental Setup

The experiments are conducted on the Lombard GRID speech corpus [19]. This dataset includes 54 speakers (30 females and 24 males) with 50 normal and Lombard parallel utterances per speaker. The utterances are meaningless sentences built as a concatenation of 6 randomly chosen words. The speech corpus is split into a training set (80%), a development set (10%) and a test set (10%) where each speaker appears proportionally. In fact, the main goal of the proposed objective evaluation is to assess the learning performance of the systems. Since each speaker may use different Lombard strategies, it is important to keep the same speakers in the training and test sets. In contrast, for the subjective evaluations, where we aim at evaluating the generalization properties of the systems, different

model	$\mathbf{f0}^{st}$		$\mathbf{e}^{dB}$		$\mathbf{mc}$
	MSE (st)	corr. (%)	MSE (dB)	corr. (%)	MCD (dB)
reference	4.34	45.3	10.03	91.3	5.53
GMM	3.34	47.7	8.37	91.3	4.65
FFNN	2.63	59.1	5.89	93.7	4.21
FR UD	2.62	59.7	5.70	94.1	4.27
FR BD	2.61	59.8	5.67	94.1	4.19
GRU UD	2.59	60.8	5.69	94.1	4.20
GRU BD	2.58	61.6	**5.61**	**94.2**	4.16
LSTM UD	2.60	60.8	5.77	93.9	4.19
LSTM BD	**2.57**	**62.2**	5.76	93.9	**4.10**

Table 2. A summary of the performance results. In this case $\mathbf{f0}^{st}$ and $\mathbf{e}^{dB}$ have been used with their raw values.

model	$\mathbf{f0}^{st}$		$\mathbf{e}^{dB}$		$\mathbf{mc}$
	MSE (st)	corr. (%)	MSE (dB)	corr. (%)	MCD (dB)
reference	4.34	45.3	10.03	91.3	5.53
GMM	3.34	49.9	8.44	91.5	4.72
FFNN	2.52	64.3	6.04	93.9	4.30
FR UD	2.54	63.6	5.88	94.1	4.31
FR BD	2.53	64.1	5.84	94.2	4.26
GRU UD	2.55	64.6	5.81	**94.3**	4.26
GRU BD	2.53	**65.4**	**5.80**	**94.3**	4.19
LSTM UD	**2.51**	**65.4**	5.82	94.2	4.29
LSTM BD	2.52	**65.4**	5.86	94.1	**4.18**

Table 3. A summary of the performance results obtained with the CWT coefficients of $\mathbf{f0}^{st}$ and $\mathbf{e}^{dB}$.

speakers are placed in the training and test sets. We here follow a leave-one speaker out strategy and thus using only 53 speakers in the training set. In any case we apply a speaker-specific zero mean and unit variance normalization to the features in order to remove speaker-specific traits and focus on the speaking style characterization. Finally, we use the delta, and delta-delta, features for every features and every model even for the RNN as preliminary tests showed that it was still beneficial for the learning.

A total of eight different configurations are trained : GMM, FFNN, FR UD/BD, GRU UD/BD and LSTM UD/BD. As we also want to study the influence of the CWT on the learning task, every configuration is trained with and without CWT.

3.2. Objective Evaluation

The converted features are compared to the original Lombard features by computing the correlation and the Root Mean Squared Error (RMSE) for $\mathbf{f0}^{st}$ and $\mathbf{e}^{dB}$, and the mean Mel-Cepstral Distortion (MCD) for $\mathbf{mc}$ to measure the spectral distortion as follows :

$$\text{MCD}(\hat{\mathbf{mc}}_i, \mathbf{mc}_i) = \frac{10}{\log 10}\sqrt{\sum_{m=1}^{24}(\hat{\mathbf{mc}}_{m,i} - \mathbf{mc}_{m,i})^2}\,. \quad (9)$$

The references correspond to the neutral speech features to provide an anchor when no conversion is applied. All the planned configurations have been trained using Hyperband and the results can be seen on table 2 with the raw features (without CWT) and on table 3 on the transformed CWT coefficients.

First, in comparing the different models with the reference (table 2), one can see that all the DNN models effectively learned the

model	$\mathbf{f0}^{st}$		$\mathbf{e}^{dB}$		$\mathbf{mc}$	
	nocwt	cwt	nocwt	cwt	nocwt	cwt
reference	-2.47	-2.47	-7.65	-7.65	-0.55	-0.55
GMM	0.10	0.55	1.24	0.91	-0.12	-0.18
FFNN	1.62	1.83	6.27	5.10	0.67	0.63
FR UD	1.56	1.73	6.75	5.83	0.70	0.58
FR BD	1.72	1.80	7.09	**5.93**	**0.71**	**0.66**
GRU UD	1.78	1.98	7.17	5.85	0.60	0.62
GRU BD	1.72	1.96	7.34	5.74	0.63	0.63
LSTM UD	1.88	1.98	7.13	5.92	0.65	0.59
LSTM BD	**2.04**	**2.02**	**7.46**	5.30	0.60	**0.66**

Table 4. Scaled net divergence $d_{net} \times 100$ for every model configurations with or without CWT. The factor of 100 is for better reading.

transformation strategies and outperform the GMM model by far. However, the difference in performance between DNN models are not statistically significant ($p > 0.1$) due to high standard deviations. Nevertheless, RNNs seems more promising even additional tests with more data would be needed to confirm this. The BD version of every RNN also slightly outperforms its UD counterpart as it allows to predict the features based on the future observations. It also uses twice as much parameters and more training data may be needed to take full advantage of these variants.

Similar observations can be made on table 3 even though the $\mathbf{e}^{dB}$ MSE and $\mathbf{mc}$ MCD are slightly worse. This performance loss is easily understandable as we are increasing the number of input and thus the number of parameters to optimize with the same number of examples. However, the $\mathbf{f0}^{st}$ feature learning task greatly benefits from the use of the CWT with a significant increase in terms of correlation. In fact, nearly all DNN models perform statistically better with, for example, a mean $\mathbf{f0}^{st}$ correlation increase of 5.3% ($p = 0.055$) for the FFNN, of 4.3% ($p = 0.034$) for the FR BD, or of 4.62% ($p = 0.026$) for the LSTM UD. The only insignificant increase comes from the LSTM BD which already had the best performance without CWT and thus less room for improvement. The CWT provides an improved learning but also reduces the gap between the models.

Figure 3 shows histogram plots of the three main timbre and prosody features i.e. $\mathbf{f0}_i$, $\mathbf{e}_i^{dB}$ and $\mathbf{mc}_{1,i}$ (spectral slope) in normal, natural Lombard and two converted Lombard styles. The two selected models for the converted Lombard style are FFNN without CWT, the current baseline model, and GRU BD with CWT, our model that objectively performed the best. We can clearly see that in both cases the feature distributions of the converted speech tend to be closer to the natural Lombard speech feature distribution. To better quantify this observation, we compute the net divergence d_{net} that measures the relative distance of the feature distributions to the natural normal and Lombard distributions. The net divergence is computed with the Jenson-Shannon divergence (JSD), a symmetrized version of the Kullback - Leibler divergence $\mathrm{D}(P||Q)$, as follows:

$$d_{net} = \mathrm{JSD}(\text{conv.}||\text{normal}) - \mathrm{JSD}(\text{conv.}||\text{Lombard})\,, \qquad (10)$$

$$\text{where} \quad \mathrm{JSD}(P||Q) = \frac{1}{2}\mathrm{D}((P||\frac{P+Q}{2}) + \frac{1}{2}\mathrm{D}((Q||\frac{P+Q}{2})\,.$$

The net divergences for all models are given in table 4. It can be seen that feature distributions of converted speech with DNN models are closer to the natural Lombard speech distribution than the neutral speech distribution ($d_{net} > 0$). The increase of performance brought by the CWT analysis is also clear here for the $\mathbf{f0}^{st}$ with a net divergence increase for all models but the LSTM BD : the

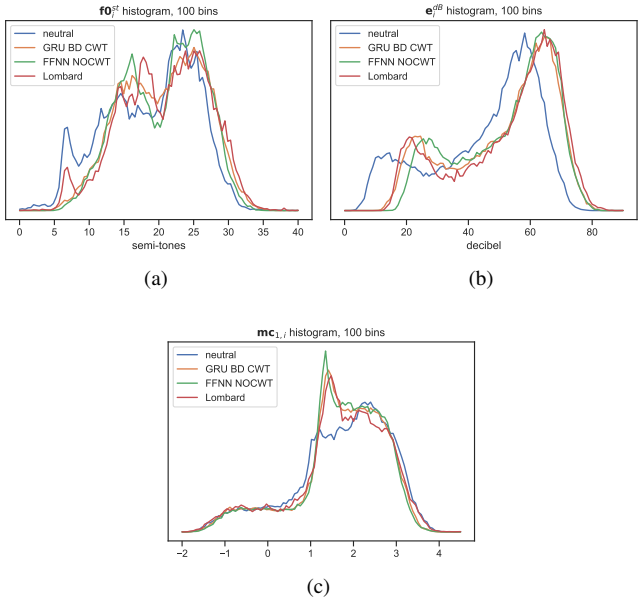

Fig. 3. Histograms of the distributions of important timbre and prosody speech features. Four distributions are displayed in each figure corresponding to the normal utterances, the Lombard style utterances and the utterances converted by two models : FFNN without CWT and GRU BD with CWT. (a) for the semi-tone scaled fundamental frequency, (b) for the dB scale energy contour and (c) for the first MFCC coefficient (spectral slope)

distributions differ even more from the neutral distribution to make progress towards the Lombard one. Surprisingly, $\mathbf{e}^{dB}$ net divergence is decreasing with CWT which seems to be contradictory with the plots. However if we compute the JSD(conv.||Lombard) alone we understand that even though it is decreasing, the converted Lombard distributions are coming closer to the natural normal one faster and decrease the net divergence as a result. Finally, the $\mathbf{mc}^{dB}$ net divergence is slightly worse like the MCD was previously, probably for the same reasons related to the increased number of parameters to optimize.

Our preliminary informal listening experiences tend to confirm the efficiency of the conversion methods, and in particular of the RNN approaches exploiting CWT transformed features. Some sound examples are given on our companion web site[1].

4. CONCLUSION

In this paper, we have proposed and evaluated several strategies for neutral to Lombard speech conversion. We have in particular shown that it is beneficial to consider a multi-resolution wavelet representation to represent the timbre and prosody speech features. The sound examples provided and the informal listening experiences done so far show that the algorithms have good generalization properties for converting speech from unknown speakers (i.e. not seen in the training phase). Future work will be dedicated to more formal listening experiences to confirm these initial results.

[1] https://perso.telecom-paris.fr/egentet/ssclombard

5. REFERENCES

[1] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "Gmm-based voice conversion applied to emotional speech synthesis," in *Eighth European Conference on Speech Communication and Technology (Eurospeech)*, 2003.

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

[3] 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, no. 5, pp. 134–138, 2012.

[4] R. Verma, P. Sarkar, and K. S. Rao, "Conversion of neutral speech to storytelling style speech," in *Eighth International Conference on Advances in Pattern Recognition (ICAPR)*. IEEE, 2015, pp. 1–6.

[5] H. Ming, D. Huang, M. Dong, H. Li, L. Xie, and S. Zhang, "Fundamental frequency modeling using wavelets for emotional voice conversion," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 804–809.

[6] H. Ming, D. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional lstm modeling of timbre and prosody for emotional voice conversion," 2016.

[7] K. I. Nordstrom, G. Tzanetakis, and P. F. Driessen, "Transforming perceived vocal effort and breathiness using adaptive pre-emphasis linear prediction," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 6, pp. 1087–1096, 2008.

[8] A. R. López, S. Seshadri, L. Juvela, O. Räsänen, and P. Alku, "Speaking style conversion from normal to lombard speech using a glottal vocoder and bayesian gmms," *Proc. Interspeech 2017*, pp. 1363–1367, 2017.

[9] S. Seshadri, L. Juvela, O. Räsänen, and P. Alku, "Vocal effort based speaking style conversion using vocoder features and parallel learning," *IEEE Access*, vol. 7, pp. 17230–17246, 2019.

[10] S. Seshadri, L. Juvela, J. Yamagishi, O. Räsänen, and P. Alku, "Cycle-consistent adversarial networks for non-parallel vocal effort based speaking style conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6835–6839.

[11] K. Nathwani, G. Richard, B. David, P. Prablanc, and V. Roussarie, "Speech intelligibility improvement in car noise environment by voice transformation," *Speech Communication*, vol. 91, pp. 17–27, 2017.

[12] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Mal. de L'Oreille et du Larynx*, pp. 101–119, 1911.

[13] D-Y. Huang, S. Rahardja, and E. P. Ong, "Lombard effect mimicking," in *Seventh ISCA Workshop on Speech Synthesis*, 2010.

[14] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," in *Acoustics, Speech and Signal Processing, ICASSP. IEEE International Conference on*. IEEE, 2008, pp. 3933–3936.

[15] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," *arXiv preprint arXiv:1603.06560*, 2016.

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

[17] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological society*, vol. 79, no. 1, pp. 61–78, 1998.

[18] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*. IEEE, 2000, vol. 3, pp. 1315–1318.

[19] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown, "A corpus of audio-visual lombard speech with frontal and profile views," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. EL523–EL529, 2018.