Whisper-to-speech conversion using restricted Boltzmann machine arrays

Jing-jie Li, Ian V. McLoughlin, Li-Rong Dai and Zhen-hua Ling

Whispers are a natural vocal communication mechanism, in which vocal cords do not vibrate normally. Lack of glottal-induced pitch leads to low energy, and an inherent noise-like spectral distribution reduces intelligibility. Much research has been devoted to processing of whispers, including conversion of whispers to speech. Unfortunately, among several approaches, the best reconstructed speech to date still contains obviously artificial muffles and suffers from an unnatural prosody. To address these issues, the novel use of multiple restricted Boltzmann machines (RBMs) is reported as a statistical conversion model between whisper and speech spectral envelopes. Moreover, the accuracy of estimated pitch is improved using machine learning techniques for pitch estimation within only voiced (V) regions. Both objective and subjective evaluations show that this new method improves the quality of whisper-reconstructed speech compared with the state-of-the-art approaches.

Introduction: Speech is a flexible communication mechanism through which people can converse using normal voice, can shout over long distances or whisper (an unvoiced (UV) mode) for private consideration communications. Whispers are often seen as a kind of degraded speech, differing through lack of voiced (V) pitch (f0) and with reduced energy. In fact, those who suffer from voice box diseases or have undergone a laryngectomy, may only be able to produce whispers. Much research effort has been spent on whisper-to-speech reconstruction [1], with Gaussian mixture model (GMM)-based voice conversion (VC) methods being state-of-the-art at present [2]. GMMs model the joint probability density of spectral parameters extracted from parallel whisper and normal speech, to subsequently transform whisper spectral parameters into those resembling speech. A second GMM models the joint probability between whisper spectrum and the f0 extracted from parallel target speech. This then generates an f0 excitation from whisper spectral parameters for reconstruction. Output speech is synthesised from the transformed spectral parameters and estimated f0. Beyond this, a further GMM is often used to reconstruct aperiodic components for additional naturalness [2]. Despite good performance, GMM SVC (statistical voice conversion) systems suffer from several issues: (i) normally they can only model high-order low-dimensionality spectral features such as mel-cepstra (i.e. spectral envelope acted on by a filterbank) [2]. However, the eventual inverse transformation from estimated mel-cepstra back into spectral envelope ignores much detail. Usually, the converted spectral parameters become over-smoothed since they are mainly determined by the weighted sum of mean vectors of each Gaussian mixture, causing a ‘muffled’ sound. This is mitigated partly by utilising dynamic spectral features, with techniques such as maximum output probability parameter generation (MOPPG) and global variance; (ii) the important V/UV decisions are derived by thresholding the f0, estimated from spectral parameters, whereas it may be better to directly classify speech as V or UV; and (iii) with the exception of [3], f0 is modelled jointly over both V and UV phonemes.

New approach: This Letter advances the state-of-the art in statistical whisper-to-speech conversion in three areas: (i) restricted Boltzmann machines (RBMs) [4] and deep learning techniques [5] are used for reconstruction. We are prompted by their success in similar tasks such as text-to-speech [6] and VC [7]; however, the main reason is that it allows us to use higher-dimensional spectral envelope information for reconstruction rather than mel-cepstra; (ii) decoupling the V/UV decision from the f0 estimation GMM, instead evaluating support vector machine (SVM) and dedicated GMM for this task; and (iii) this allows V and UV phonemes to be modelled in different feature spaces, and the estimated f0 derived only from V phonemes [3].

In detail, spectral features are represented as envelopes (not high-order mel-cepstra) extracted following [8] for whispers and parallel speech, synchronised by dynamic time warping (DTW), shown in Fig. 1. The joint spectral density space is modelled using multiple RBMs (instead of a single GMM), which has been shown by other authors to better match inter-speaker spectral correlation [7]. V/UV decisions are made on input whispers using a SVM to divide spectral features into V and UV. f0 estimation is then performed specifically on V frames using GMM or support vector regression (SVR) evaluated by comparing reconstructed speech that of the baseline GMM method of [2].

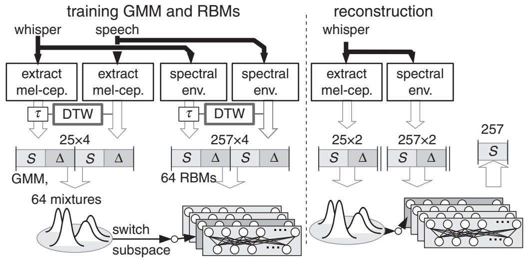

Fig. 1 *Training and operating phases for spectral reconstruction*

RBM-based spectral conversion: RBMs are bipartite undirected graph models where visible units $\boldsymbol{v} = [v_1, \ldots, v_V]^\top$ are connected to hidden units $\boldsymbol{h} = [h_1, \ldots, h_H]^\top$ by weight matrix $\boldsymbol{W}_{V\times H}$. V and H denote the number of visible and hidden layer units. Given input $\boldsymbol{v}$ to the visible units, the energy function of a Gaussian RBM is defined as

$$E(\boldsymbol{v}, \boldsymbol{h}; \theta) = \frac{1}{2}(\boldsymbol{v} - \boldsymbol{a})^\top(\boldsymbol{v} - \boldsymbol{a}) - \boldsymbol{b}^\top \boldsymbol{h} - \boldsymbol{v}^\top \boldsymbol{W}\boldsymbol{h} \tag{1}$$

where $\theta = \{\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}\}$ are the model parameters, $\boldsymbol{a} = [a_i, \ldots, a_V]^\top$ and $\boldsymbol{b} = [b_i, \ldots, b_H]^\top$ are the bias of visible and hidden units. The joint probability distribution function (PDF) is then defined as

$$P(\boldsymbol{v}) = \frac{1}{\mathcal{Z}}\sum_h \exp(-E(\boldsymbol{v}, \boldsymbol{h}; \theta)) \tag{2}$$

where $\mathcal{Z} = \int_{\boldsymbol{v}} \sum_h \exp(-E(\boldsymbol{v}, \boldsymbol{h}; \theta))\,\mathrm{d}\boldsymbol{v}$ is the partition function. The RBM parameters $\theta = \{W, \boldsymbol{a}, \boldsymbol{b}\}$ are obtained via the contrastive divergence (CD) algorithm with a maximum-likelihood criteria.

During conversion, the converted spectral feature vector $\boldsymbol{y}_t^*$ is obtained by maximising the conditional probability given input vector $\boldsymbol{x}_t$

$$\boldsymbol{y}_t^* = \arg\max_{\boldsymbol{y}_t} P(\boldsymbol{y}_t|\boldsymbol{x}_t, \theta) \tag{3}$$

Toda *et al.* [2] demonstrated that the conditional probability can be approximated without obvious performance loss by

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, \theta) \simeq P(\boldsymbol{y}_t|\boldsymbol{x}_t, m^*, \theta) = \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{m^*,t}, \boldsymbol{\Sigma}_{m^*,y}) \tag{4}$$

where m^* is the optimum subspace that has biggest posterior probability of the given input feature vector, $\mathcal{N}$ denotes Gaussian PDF, $\boldsymbol{\Sigma}_{m^*,y}$ is the diagonal covariance matrix of the target normal speech in the m^*th spectral feature subspace and $\boldsymbol{\mu}_{m^*,t}$ is the mode of the m^*th RBM, which can be obtained through solving the following optimisation:

$$\boldsymbol{\mu}_{m^*,t} = \arg\max_{\boldsymbol{y}_t} P(\boldsymbol{v}_t|m^*, \theta) \tag{5}$$

$\boldsymbol{v}_t = [\boldsymbol{x}_t, \boldsymbol{y}_t]$ are visible unit inputs to the m^*th RBM. With no closed-form solution, it can be solved by gradient descent, given learning rate α:

$$\boldsymbol{\mu}^{\text{new}}_{m^*,t} = \boldsymbol{\mu}^{\text{old}}_{m^*,t} + \alpha\frac{\partial \log P(\boldsymbol{v}_t)}{\partial \boldsymbol{y}_t} \tag{6}$$

$$\frac{\partial \log P(\boldsymbol{v}_t)}{\partial \boldsymbol{y}_t} = -(\boldsymbol{y}_t - a^{(y)}) + \sum_{j=1}^{H} \sigma(b_j + v_t^\top)\boldsymbol{w}_j^{(y)} \tag{7}$$

and $\sigma(x) = (1 + \mathrm{e}^{-x})^{-1}$ is the logistic sigmoid function. According to the characteristic Gaussian distribution, the conditional probability in (4) is maximised when $\boldsymbol{y}_{m^*,t} = \boldsymbol{\mu}_{m^*,t}$. In (5) and (6), note that the converted spectral feature is not dominated by the mean of m^*th target spectral feature subspace (as it would be in a GMM system). Moreover, this allows us to implement RBM-based conversion by modifying a baseline GMM [2]. Enhancements developed for GMM-based systems, such as dynamic features and MOPPG, are still compatible with the proposed architecture. The proposed system comprises spectral envelope (Fig. 1) and f0 conversion modules. During training, a V/UV decision model (e.g. GMM or SVM) is first trained using the mel-cepstra static (s) and dynamic (Δ) features of whispers with V/UV data from DTW-aligned normal speech (from extracted f0 tracks). Next, an f0 estimation model is trained for the V subspace only using whisper spectral features and the extracted speech f0. Meanwhile, multiple RBMs are trained using spectral envelope features from the V subspace regions to model the joint spectral density between whispers and time-aligned speech shown in (2). Reconstruction begins with a frame-wise V/UV decision from input whispers. For V frames, f0 is estimated, and spectral envelope features are obtained from the RBMs using (5) and the MOPPG algorithm. UV output uses amplitude-normalised whispered frames.

Evaluation: The proposed methods were evaluated as follows: 25-order mel-cepstra and 257-order spectral envelopes were extracted from whispers and corresponding speech [8]. DTW was computed between the whisper and speech mel-cepstra (and used for mel-cepstra, spectral envelopes, V/UV regions and f0 from the parallel training data). Parallel whisper and speech recordings from a whispered TIMIT database (wTIMIT) [9] (female speaker 002 and male speaker 003) were divided into test data of 10 000 analysis frames, training data (~180 000 and frames). We first assess V/UV decision accuracy. GMM and SVM methods were evaluated in terms of error rates for different lengths of concatenated GMM input vectors, and for SVM in Table 1. Principal component analysis was used to reduce the high-dimensional features to a 50-dimension (50D) vector. Table 1 reveals that the optimal context size is ±5 frames, and that the SVM error rate slightly exceeds that of the optimal GMM choice. Overall, these methods contribute a V/UV error rate of about 9% to the subsequent spectral modelling of V frames, comparable with 9.76% in [3].

Table 1: V/UV error rates of GMMs for SVM and various GMMs

	Static (%)	±1 (%)	±3 (%)	±5 (%)	±7 (%)	SVM ± 5 (%)
V → U	7.3	5.1	5.28	5.09	5.55	4.39
U → V	6.58	4.95	4.41	3.77	3.54	5.08
Total	13.88	10.05	9.69	8.86	9.09	9.47

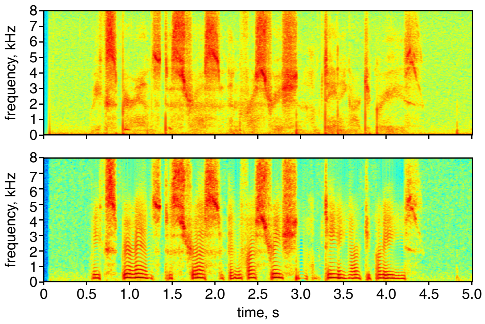

Fig. 2 *Whisper (top) and reconstructed (bottom) spectrograms*

Table 2: f0 estimation for different regression models

	Baseline GMM	V-only GMM	V-only SVR
RMSE (Hz)	29.95	12.97	13.80
Correl. coeff.	0.26	0.61	0.49

Table 3: MOS and subjective preference scores of GMM and RBM

	Mean (95% confidence)		Preference (remainder indicates none)	
	Female speech	Male speech	Female speech (%)	Male speech (%)
GMM	2.25 (±0.18)	2.35 (±0.13)	2.5	2.7
RBM	2.91 (±0.16)	2.87 (±0.15)	73.8	54

Secondly, f0 estimation accuracy is compared for different regression models. Table 2 gives the root mean squared error (RMSE) and correlation coefficient. Evidently, a significant performance gain is achieved by separately modelling the V and UV subspaces (i.e. estimate f0 from V frames only), with SVR achieving similar performance. Finally, the proposed multiple-RBM reconstruction system was evaluated against the baseline [2] with 64 mixtures. The baseline GMM was then used to divide the analysis frames into 64 spectral subspaces. One RBM, with 1028 visible and 100 hidden units, was trained per subspace using the CD algorithm [5]. Both static and dynamic spectral envelope features were used, and MOPPG employed to generate final static features for re-synthesis. f0 was estimated as described above for GMM-classified V frames only. For subjective evaluation, eight students with no known hearing impairments assessed whispers from reconstructed baseline and proposed methods in a soundproofed room, wearing headphones. Testing used a mean opinion score (MOS) protocol with 50 sentences per condition. A separate two-alternative preference test was also conducted. The results, shown in Table 3, clearly indicate that the RBM method achieves higher MOS and is the clearly preferred method. The proposed RBM system achieves a log spectral distortion (LSD) of 6.10 (±0.15), compared with 5.96 (±0.13) for the 64-mixture GMM baseline and 11.07 (±0.29) for the whispers. In general, the nonlinearity of (6) coupled with the avoidance of a mel-cepstral transformation loss improves the fidelity of modelled fine detail. Fig. 2 shows an example spectrogram.

Conclusion: This Letter has proposed and evaluated three improvements to state-of-the-art GMM-based whisper-to-speech reconstruction systems: (i) decoupling the V/UV decision from f0 estimation, potentially allowing better performance for both tasks; (ii) modelling f0 for V subspaces only achieved a significant improvement over the usual method of modelling f0 for combined V and UV subspaces; and (iii) the first application of multiple RBMs for whisper-to-speech VC. RBMs allowed higher-dimensional spectral envelope features to be used: a 1028D GMM would be extremely difficult to train directly. Results indicate a very strong preference for the RBM-reconstructed speech, as well as improved MOS over the GMM system.

30 May 2014
doi: 10.1049/el.2014.1645
One or more of the Figures in this Letter are available in colour online.

Jing-jie Li, Ian V. McLoughlin, Li-Rong Dai and Zhen-hua Ling (*The University of Science and Technology of China, Hefei, Anhui, People's Republic of China*)

E-mail: ivm@ustc.edu.cn

References

1 Sharifzadeh, H.R., McLoughlin, I.V., and Ahmadi, F.: 'Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec', *IEEE Trans. Biomed. Eng.*, 2010, **57**, (10), pp. 2448–2458
2 Toda, T., Nakagiri, M., and Shikano, K.: 'Statistical voice conversion techniques for body-conducted unvoiced speech enhancement', *IEEE Trans. Audio Speech Lang. Process.*, 2012, **20**, (9), pp. 2505–2517
3 Tran, V.-A., Bailly, G., Loevenbruck, H., and Toda, T.: 'Improvement to a NAM-captured whisper-to-speech system', *Speech Commun.*, 2010, **52**, (4), pp. 314–326
4 Hinton, G.E., and Salakhutdinov, R.: 'Reducing the dimensionality of data with neural networks', *Science*, 2006, **313**, (5786), pp. 504–507
5 Hinton, G.E.: 'Training products of experts by minimizing contrastive divergence', *Neural Comput.*, 2002, **14**, (8), pp. 1771–1800
6 Ling, Z.-H., Deng, L., and Yu, D.: 'Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis', *IEEE Trans. Audio, Speech, Lang. Process.*, 2013, **21**, (10), pp. 2129–2139
7 Chen, L.-H., Ling, Z.-H., Song, Y., and Dai, L.-R.: 'Joint spectral distribution modeling using restricted Boltzmann machines for voice conversion'. Proc. INTERSPEECH, Lyon, France, September 2013, pp. 3052–3056
8 Kawahara, H., and Irino, T.: 'Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation', in Divenyi, P. (Eds.): 'Speech Separation by Humans and Machines' (Springer, 2005), pp. 167–180
9 Lim, B.P.: 'Computational differences between whispered and non-whispered speech'. PhD thesis, UIUC, 2010