Citation for published version


DOI

https://doi.org/10.1049/el.2014.1645

Link to record in KAR

http://kar.kent.ac.uk/48793/

Document Version

UNSPECIFIED

Copyright & reuse
Content in the Kent Academic Repository is made available for research purposes. Unless otherwise stated all content is protected by copyright and in the absence of an open licence (eg Creative Commons), permissions for further reuse of content should be sought from the publisher, author or other copyright holder.

Versions of research
The version in the Kent Academic Repository may differ from the final published version. Users are advised to check http://kar.kent.ac.uk for the status of the paper. Users should always cite the published version of record.

Enquiries
For any further enquiries regarding the licence status of this document, please contact: researchsupport@kent.ac.uk

If you believe this document infringes copyright then please contact the KAR admin team with the take-down information provided at http://kar.kent.ac.uk/contact.html
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 formal language, voice conversing over long distances or whisper (an unvoiced (UV) mode) for private consideration communications. Whispers are often seen as a kind of degraded speech, differing from lack of voice pitch at rest (0) and with reduced energy. In fact, 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 speech [1]. With a 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 speech, synchronised by dynamic time warping (DTW), shown in Fig. 1. The joint spectral density space is modelled using multiple restricted Boltzmann machines (RBMs) [4] and deep learning techniques [5] are used for pitch estimation within only voiced (V) frames using GMM or support vector regression (SVR) evaluated by comparing reconstructed speech that of the baseline GMM method of [2].

RBM-based spectral conversion: RBMs are bipartite undirected graph models where visible units \( v = [v_1, \ldots, v_M]^T \) are connected to hidden units \( h = [h_1, \ldots, h_N]^T \). With weight matrix \( W_{v \rightarrow h} \), \( V \) and \( H \) denote the number of visible and hidden units. Given input \( v \) to the visible units, the energy function of a Gaussian RBM is defined as

\[
E(v, h; \theta) = \frac{1}{2} (v - a)^T (v - a) - b^T h - v^T Wh
\]

where \( \theta = \{W, a, b\} \) are the model parameters, \( a = [a_1, \ldots, a_M]^T \) and \( b = [b_1, \ldots, b_N]^T \) are the bias of visible and hidden units. The joint probability distribution function (PDF) is then defined as

\[
P(v) = \frac{1}{Z} \exp(-E(v, h; \theta))
\]

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

During conversion, the converted spectral feature vector \( v_\ast \) is obtained by maximising the conditional probability given input vector \( x \).

\[
y_\ast = \arg\max P(y|x, \theta)
\]

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

\[
P(y|x, \theta) \approx P(y|x, m^*, \theta) = \mathcal{N}(y|m^*, \mu_{V^\ast}, \Sigma_{V^\ast})
\]

where \( m^* \) is the optimum subspace that has biggest posterior probability of the given input feature vector, \( \mathcal{N} \) denotes Gaussian PDF, \( \Sigma_{V^\ast} \) is the diagonal covariance matrix of the target normal speech in the \( m^\ast \) th spectral feature subspace and \( \mu_{V^\ast} \) is the mode of the \( m^\ast \) th RBM. However, the estimated \( m^\ast \) by the proposed model is not dominated by the mean of \( m^\ast \) 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 \( f_0 \) conversion modules. During training, a VUV decision model (e.g. GMM or SVM) is first trained using the mel-cepstra static and dynamic (\( \Delta \) features) of whispers with VUV data from DTW-aligned normal speech (from extracted \( f_0 \) tracks). Next, an \( f_0 \) estimation model is trained for the V subspace only using whisper spectral features into V and UV. \( f_0 \) 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].
features and the extracted speech $f_0$. 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, $f_0$ 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 $f_0$ 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

<table>
<thead>
<tr>
<th>Static (%)</th>
<th>+1 (%)</th>
<th>+3 (%)</th>
<th>+5 (%)</th>
<th>+7 (%)</th>
<th>SVM ±5 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V → U</td>
<td>7.3</td>
<td>5.1</td>
<td>5.28</td>
<td>5.09</td>
<td>5.55</td>
</tr>
<tr>
<td>U → V</td>
<td>6.58</td>
<td>4.95</td>
<td>4.41</td>
<td>3.77</td>
<td>3.54</td>
</tr>
<tr>
<td>Total</td>
<td>13.88</td>
<td>10.05</td>
<td>9.69</td>
<td>8.86</td>
<td>9.09</td>
</tr>
</tbody>
</table>

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

### Table 2: $f_0$ estimation for different regression models

<table>
<thead>
<tr>
<th></th>
<th>Baseline GMM</th>
<th>V-only GMM</th>
<th>V-only SVR</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE (Hz)</td>
<td>29.95</td>
<td>12.97</td>
<td>13.80</td>
</tr>
<tr>
<td>Correl. coeff.</td>
<td>0.26</td>
<td>0.61</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Secondly, $f_0$ 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 $f_0$ 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. $f_0$ 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 $f_0$ estimation, potentially allowing better performance for both tasks; (ii) modelling $f_0$ for V subspaces only achieved a significant improvement over the usual method of modelling $f_0$ 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.

© The Institution of Engineering and Technology 2014
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