Speech reconstruction using a deep partially supervised neural network

Ian McLoughlin[1,2] ✉, *Jingjie Li*[2], *Yan Song*[2], *Hamid R. Sharifzadeh*[3]

[1]*School of Computing, The University of Kent, Medway, UK*
[2]*National Engineering Laboratory of Speech and Language Information Processing, The University of Science and Technology of China, Hefei, Anhui, People's Republic of China*
[3]*Signal Processing Laboratory, Unitec Institute of Technology, Auckland, New Zealand*
✉ *E-mail: ivm@kent.ac.uk*

Published in Healthcare Technology Letters; Received on 4th December 2016; Revised on 16th April 2017; Accepted on 1st May 2017

Statistical speech reconstruction for larynx-related dysphonia has achieved good performance using Gaussian mixture models and, more recently, restricted Boltzmann machine arrays; however, deep neural network (DNN)-based systems have been hampered by the limited amount of training data available from individual voice-loss patients. The authors propose a novel DNN structure that allows a partially supervised training approach on spectral features from smaller data sets, yielding very good results compared with the current state-of-the-art.

1. Introduction: In this Letter, larynx-related dysphonia refers to those who have undergone a partial laryngectomy or who have larynx damage, impaired brain function or nerve lesion (e.g. laryngeal palsy), or are avoiding phoned speech due to a prescribed period of voice rest [1, 2]. This leads them to produce medical whispers – a subset of whispering for medical reasons – where the vocal cords do not vibrate as they would for spoken utterances, even when producing vowels and other phonemes that are normally voiced. Such whispers exhibit reduced energy compared with speech and are easily obscured by background noise, thus the need for speech reconstruction. Simple prosthetic aids have been available since the invention of the robotic sounding electrolarynx in the 1920s, but much recent research has focused on two approaches to whisper-to-speech conversion (WSC) [1], namely codec-based and statistical voice conversion (SVC) methods. The former uses parametric conversion frameworks which decompose whisper input and then reconstruct into normal speech without model training or use of a priori information. These include the code excited linear prediction and mixed excitation linear prediction-based reconstruction technique [2, 3] plus direct conversion methods [1, 4].

Codec-based systems are fast, efficient, and simple to set up, but exhibit unnatural pitch, and mean opinion scores seldom exceed 3.5. SVC-based methods using Gaussian mixture models (GMMs) [5] reconstruct much more normal sounding speech from whispers using a priori data to build joint whisper–speech models. However, current systems are limited in only modelling compressed mel-cepstral coefficients, failing to model inter-dimensional correlation due to the restriction of a diagonal covariance matrix when the training data is limited [6]. As a result, speech converted from whispers using a GMM usually sounds 'muffled' with unusual pitch contours.

The authors previously introduced a WSC framework using arrays of restricted Boltzmann machines (RBMs) [6] acting on spectral envelope information, allowing much higher-dimensional spectral information than GMM methods. This also modelled inter-dimensional spectral correlation (due to full connectivity between hidden and visible layers). It improved on baseline GMM-based system in terms of both intelligibility and naturalness [6]; however, its architecture was significantly more complex than GMM systems, and much slower since it had to convert each frame individually using a gradient descent algorithm.

This Letter presents a new approach to improve quality while reducing complexity, as shown in Fig. 1. Rather than operate multiple RBM arrays to learn spectral feature mappings, we propose a single deep neural network (DNN) structure. Unsupervised training is used first to create two separate Gaussian–Bernoulli RBMs: one for whisper spectral features and one for frame-aligned speech spectral features, effectively performing feature coding. These trained RBMs are then stacked back-to-back – a two-layer Bernoulli–Bernoulli neural network is sandwiched between the RBMs to form a DNN stack. The middle layers are then trained in a supervised fashion to form a fully connected mapping from whisper features to frame-aligned speech features.

2. Background: GMM-based WSC methods were pioneered by Toda *et al.* [5] to convert non-audible murmur signals into speech with more normal sounding characteristics. The authors subsequently extended their techniques to convert whisper-like speech from post-partial laryngectomy patients, transforming acoustic features of whispers into normal sounding speech after being suitably trained with parallel utterance data (i.e. spoken and whispered versions of the same speech). State-of-the-art variants [7] made use of up to three GMMs: one to convert the source spectral features into target spectral features, another to convert the same source spectral features into a pitch excitation, and the last one to generate additional aperiodic components that are found in the target speech (which can enhance naturalness). Converted speech is typically synthesised by speech transformation and representation based on adaptive interpolation of weighted spectrogram (STRAIGHT) [8], using estimated spectral features, pitch, and aperiodic components. Quality tends to be good though these methods suffer from over-smoothing of detailed characteristics in the reconstructed spectra, leading to muffled speech. Unnatural prosody can arise too, due to the difficulty in estimating $f0$ from the whisper spectrum. Apart from these disadvantages, complexity is high due to the three GMMs needed, and because systems require STRAIGHT for re-synthesis.

In a recent system developed by Li *et al.* [6], the over-smoothing effect was mitigated through the use of RBM arrays which could model a much higher resolution spectral envelope than commonly achieved by the mel-cepstra of GMM systems. The RBM system also decoupled the voiced/unvoiced classification from the pitch estimation, allowing for smoother pitch reconstruction. This performance gain, however, came at a substantial computational cost.

3. Motivation for the proposed technique: The motivation to utilise a DNN architecture for the WSC task is simple: DNNs have shown their ability to infer discriminative features in a

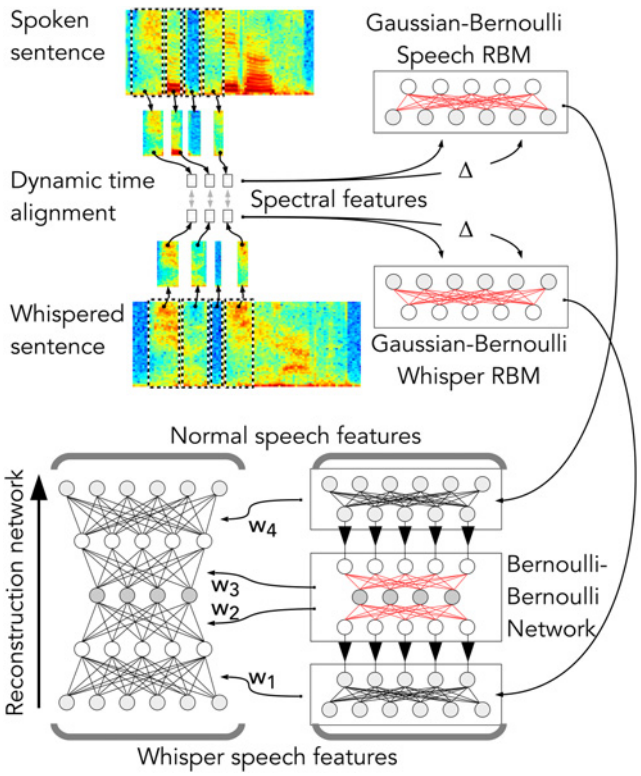

Fig. 1 *RBM networks first trained on time-aligned whisper and speech spectral features (top), then used to train feature mapping network weights (bottom)*

number of related domains including automatic speech recognition [9], language identification [10], and machine hearing [11]. They have also been applied to the field of SVC [12] as well as speech enhancement [13]. Therefore, we began to train DNN-based WSC regression models using minimum mean squared error objective criteria on spectral envelope features. We followed standard RBM-based pre-training methods using the contrastive divergence (CD) algorithm and back-propagation (BP) error-based fine-tuning. The results, however, were systems prone to over-fitting due to the limited parallel data from each speaker [Parallel training data means high-quality recordings of a patient speaking sentences with a normal voice and speaking the same sentences after laryngectomy. This is difficult to obtain in practise from patients who have already lost the ability to speak naturally.]. DNN performance was disappointing, as results will indicate in Section 9.

Without obtaining significantly more training data, one potential improvement would be to use supervised training, which is more effective than unsupervised training. Unfortunately supervision requires labelled data, which we do not have. However, we propose a novel semi-supervised DNN (semi-DNN) architecture in this Letter which first uses unsupervised training to perform feature coding, and then uses that coding to enable supervised training. Specifically, we begin by training two separate RBMs on whisper and voiced speech spectral envelopes. These are trained in an unsupervised fashion, and can thus use large-scale databases from many speakers. The RBMs output identically sized sets of binary features from spectral envelope inputs. Frame-aligned binary features from each RBM are then used as inputs to train a fully connected mapping network, in a supervised fashion, that effectively translates the RBM-extracted Bernoulli feature spaces between the two speech modes. We will see that this method reconstructs much better speech than a single DNN of equivalent size, trained in a wholly unsupervised fashion using the same data. It also outperforms existing GMM and RBM techniques.

4. WSC using standard DNN: The most direct application of DNN for WSC is a network where the input layer data is original whisper features, and the output layer data maps to parallel normal speech features. We implemented such a system, trained using parallel frame-aligned spectral envelope data. To reduce the over-smoothing of reconstructed spectra, dynamic features, and maximum output probability parameter generation (MOPPG) algorithms were deployed. These have been demonstrated in both GMM and multiple RBM-based WSC tasks to be effective at reducing over-smoothing [14].

Assume that $\boldsymbol{X}_t = [\boldsymbol{x}_t, \Delta\boldsymbol{x}_t]$ are the static and dynamic features of whispers, while the corresponding parallel features from normal speech are $\boldsymbol{Y}_t = [\boldsymbol{y}_t, \Delta\boldsymbol{y}_t]$. When training the DNN using error BP through minimum mean square error criteria, the objective function for system θ is

$$J(\theta) = -\frac{1}{2T}\sum_{t=1}^{T}\sum_{k=1}^{D}\{\hat{\boldsymbol{Y}}_{tk}(\theta) - \boldsymbol{Y}_{tk}\}^2 \tag{1}$$

The output layer of the DNN has $\boldsymbol{D}$ nodes, computed in a forward layer-wise direction. During WSC, the output from time 0 to T would be $\hat{\boldsymbol{Y}} = \{\hat{\boldsymbol{Y}}_t\}$ where $1 \le t \le T$. According to the MOPPG algorithm, the converted static feature $\hat{\boldsymbol{y}}$ is computed as follows:

$$\hat{\boldsymbol{y}} = (\boldsymbol{C}^{\mathrm{T}}\boldsymbol{D}^{(Y)}\boldsymbol{C})^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{D}^{(Y)^{-1}}\hat{\boldsymbol{Y}} \tag{2}$$

where $\boldsymbol{C}$ is a transformation matrix that maps static features into static and dynamic combined features, and $\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{y}$. While $\boldsymbol{D}^{(Y)}$ can be estimated from

$$\boldsymbol{D}^{(Y)} = \Sigma^{(YY)} - \boldsymbol{\Sigma}^{(YX)}\Sigma^{(XX)^{-1}}\Sigma^{(XY)} \simeq \Sigma_m^{(YY)} \tag{3}$$

since the elements of covariance matrix $\boldsymbol{\Sigma}^{(YX)}$ are close to zero. The DNN implementation is shown in Fig. 2.

The application of DNNs to whisper-to-speech reconstruction can simplify the training and operating process compared with the use of multiple RBMs [6]. However, in practise the DNNs trained from the same data set as the GMM and multiple RBM systems are prone to over-fitting due to insufficient training data. This will be reflected in the results reported later in Section 8.

5. Training data: For WSC tasks, training data consists of parallel sentences of whisper and target speech. Thus, models are trained with the same sentences whispered as well as spoken, with the aim of the training being to convert whispers into target speech. Systems are speaker-dependent and trained for one user at a time and sufficient good quality parallel whisper–speech data must be prepared for each user, and then aligned at a frame-level. Unfortunately, lack of training data is endemic for WSC systems, particularly as the primary target users are patients with larynx damage who can no longer produce normal speech on demand. Contrast this with the fact that DNNs require a large amount of training data in order to learn the numerous internal parameters of deep fully interconnected layers. The apparent mismatch between amount of training data and DNN requirements motivates the novel semi-DNN proposed in the next section.

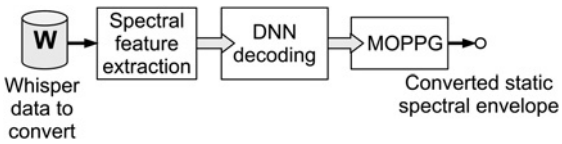

Fig. 2 *Standard DNN implementation for performing WSC*

6. Proposed semi-DNN: The core idea of the proposed semi-DNN combines unsupervised and supervised training methods together within a single composite deep network. With reference to Fig. 1, the first stage is to train two separate Gaussian–Bernoulli RBMs on whisper and corresponding parallel (frame-aligned) speech features using the unsupervised CD algorithm [12]. In practise, each RBM can model their respective feature space quite well, mapping from Gaussian to Bernoulli space. The number of nodes in the hidden layer is much lower than the number of nodes in the visible layers, which effectively performs a feature coding or information compression procedure. This aspect is crucial in reducing the subsequent training requirements of the supervised layer which will be added later. Coding implies that input data can be reconstructed from the corresponding hidden data, with the degree of reconstruction precision being dependent on both the degree of redundancy in the training data and the dimensionality of both layers. Assuming, as before, that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are whisper and speech spectral data (both static and dynamic spectral envelope features) from time 1 to time T, time aligned on a frame-by-frame basis. The hidden nodes $\hat{\boldsymbol{h}}_x$ corresponding to $\boldsymbol{X}$ can be computed as follows:

$$\hat{\boldsymbol{h}}_{tk}^{x} = \begin{cases} 1, & P(\boldsymbol{h}_{tk}^{x} = 1|\boldsymbol{X}_t,\ \theta_X) \geq 0.5 \\ 0, & P(\boldsymbol{h}_{tk}^{x} = 1|\boldsymbol{X}_t,\ \theta_X) < 0.5 \end{cases}, \quad \text{for} \quad 1 \leq t \leq T \qquad (4)$$

where $P(\boldsymbol{h}_{tk}^{x} = 1|\boldsymbol{X}_t,\ \theta_X) = \mathrm{Ber}\left(\boldsymbol{h}_{tk}^{x}|\int(W_{tk}^{\mathrm{T}}\boldsymbol{X}_t + c)\right)$ and $\int(\boldsymbol{x}) = 1/(1+\exp(-\boldsymbol{x}))$ is the logistic function. The hidden data $\hat{h}_y$ can be obtained in a similar way. After $\hat{\boldsymbol{h}}_x$ and $\hat{\boldsymbol{h}}_y$ are obtained, they are subsequently used as training data for the hidden middle layers of the mapping DNN (the meat in the sandwich), trained in a supervised fashion using the BP algorithm as illustrated in Fig. 3. Finally, the resulting semi-DNN can be obtained by stacking the two unsupervised RBMs and the middle network together as shown on the right-hand side in Fig. 1.

This semi-supervised training method not only significantly reduces the number of parameters that need to be trained by the BP algorithm, but also stabilises the parameters of the middle hidden layers compared with standard DNN training methods, in part due to the already coded Bernoulli input which is thought to have the effect of separating the mapping and feature extraction functions of the network.

7. Semi-DNN training: In detail, training first involves spectral features being extracted from corresponding whispered and spoken utterances. Second, the features are then aligned using dynamic time warping (DTW), since though the utterances were spoken by the same person, there are significant timing differences between whispering and speaking [15], and these must be corrected to provide frame-level feature alignment. Third, two RBMs are trained separately using normalised whisper spectral data and normalised speech spectral data, both in an unsupervised fashion. Fourth, the middle network in the DNN is trained with data passed through the two RBMs. This is supervised training. Finally, the two RBMs and middle mapping network are concatenated together to form a complete DNN. During WSC operation, spectral feature vectors consisting of concatenated static and dynamic spectral envelope features are extracted from whispers, input to the DNN, and an output obtained by computing layer-wise in an upward direction with respect to the bottom right network in Fig. 1. Finally, converted static spectral envelopes are obtained using the MOPPG algorithm, and the reconstructed speech is synthesised from the static spectral envelope combined with an estimated $f0$ obtained using the same method as in [6].

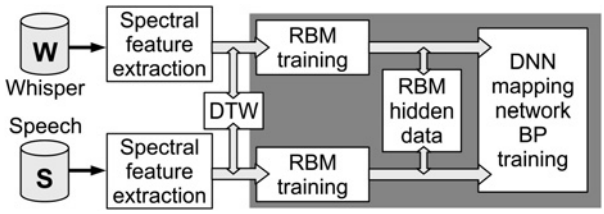

Fig. 3 *Semi-DNN training methodology showing the two-pass training arrangement in the shaded box to the right*

8. Performance evaluation: The effectiveness of the system is evaluated for speaker-dependent WSC using both objective and subjective criteria with results compared with a baseline GMM system [5], the more recent RBM array structure [6], and the direct DNN implementation from Section 4. For fairness, each system makes use of identical pitch contour reconstruction data, and thus the performance comparison is based on spectral feature reconstruction fidelity.

Parallel speech and whisper data are obtained from the whisper-TIMIT (wTIMIT) open-source whisper corpus on a per-speaker basis. Testing data comprises 15 utterances selected randomly from the 450 utterances of a single speaker in wTIMIT, with the remaining 435 utterances used as a training set (~15 min worth). Frame size is 40 ms, with an overlap of 35 ms, and 513 log spectral envelope parameters per frame (i.e. DC and 512 frequency bins). In addition, 25-order mel-cepstra are simultaneously extracted from each frame – these are used for both objective scoring as well as for the DTW alignment (since the spectral envelope dimensionality is so large it makes DTW almost unworkable).

The standard DNN implementation has input and output layers with 1026 nodes comprising 513 spectral features and 513 differential features. The network contains two hidden layers of dimension 1024. When training the DNN in its supervised fine-tune stage through the BP algorithm, the learning rate is 0.5. A weight decay strategy is not used for this system.

During the RBM-based layer-wise pre-training, the batch size is set to 10, with a single Gibbs sampling step, and learning rate of 0.0001. RBM training is iterated 100 times with a momentum term of 0.5 for the first five iterations and 0.9 for the following 95 iterations.

The proposed semi-DNN input and output layers have identical dimensions to the standard DNN system, namely 1026. It comprises three hidden layers with dimensions of 1024, 512, and 1024. This has been chosen to ensure that the number of trainable parameters in the proposed semi-DNN match those of the standard DNN. Additionally, the training configuration for the two RBMs and middle mapping network in the semi-DNN is exactly the same as that used in the standard DNN. Thus the training data, training settings, number of training parameters, and feature dimensionality is the same between the two systems. Any difference in performance is thus achieved solely by the novel mapping structure and partially supervised training that this structure makes possible.

The detailed configuration of the GMM and multiple RBM array WSC systems used for comparison are as described in [6], but in this case will use the same training and evaluation data set as the DNNs above.

In the subjective and objective evaluations that follow, there are thus six possible items of comparison: input whispers, corresponding speech, and reconstructed speech from GMM, RBM, DNN, and semi-DNN. Each of these six can thus be compared for all utterances.

Subjective evaluation: Six naive student volunteers with normal hearing participated in binary preference listening tests to evaluate the four WSC models. In each sitting, every listener was presented with a set of randomly sequenced pairs of recordings and asked to state their preference between the two recordings, or to indicate 'no preference'. The recordings were sentences reconstructed from the four systems under evaluation. Four binary tests conducted in this way per sentence can, therefore, subjectively discriminate

between each of the four models. This was repeated for all evaluation sentences for each listener.

Objective evaluation: For objective scoring, cepstral distortion is used to evaluate the spectral distance between whispers, the converted speech from different models, and the parallel normal speech. The DTW algorithm is used to align 25-order mel-cepstra of source speech and target speech, and hence provide a frame-level alignment for parallel comparison. During testing, the mel-cepstra of reconstructed speech from the RBM, standard DNN and proposed semi-DNN models were generated directly from the reconstructed spectral envelope output.

To explore further, three other objective methods were also applied; symmetrical Itakura–Saito (IS) distance (i.e. the average IS distance in both directions), log-likelihood ratio (LLR), and segmental SNR [15]. These were computed between each of the DTW-aligned reconstructed outputs from the four models plus the whisper input, and the corresponding speech recording.

9. Results: The subjective evaluation results are reported in Table 1. Each row records the mean two-way preference of the six listeners on each of 15 utterances. Each column identifies the proportion of preferences reported for a particular WSC method, apart from the last column which indicates where listeners were unable or unwilling to express a preference.

Beginning with the top of the table, the first three rows indicate a very significant preference for the proposed semi-DNN over either GMM, DNN, or multiple RBMs. In particular, comparing rows 1 and 4 the preference for GMM is reduced from 43 to 13% through the adoption of the novel semi-supervised training architecture. This result very clearly demonstrates the effectiveness of the proposed technique.

Cepstral distortion scores, reported in Table 2, indicate that all four models improve on the correspondence of the whisper input to matching speech in a spectral distance sense. However, the semi-DNN and RBM models do not score as well as the GMM and DNN reconstructions, in marked contrast to the subjective results. This matches the findings of the RBM-GMM evaluation in [6] and highlights a difference between objective and subjective evaluations. To explore further, Fig. 4 plots the mean IS, LLR, and segmental SNR scores – these are objective measures of the similarity of the given signal to the corresponding speech (smaller meaning more similar). Apart from segmental SNR, all methods improve on the whispers, with the proposed semi-DNN method performing well in each evaluation.

10. Further analysis: To provide additional insight into these systems, Fig. 5 plots the linear predictive coefficients-derived spectral envelope for an example utterance (0.5 s of voiced speech and the corresponding time-aligned sections from the whisper and reconstructed outputs).

Table 1 Results of the four binary subjective evaluation tests

GMM	RBM	DNN	Semi-DNN	No preference
13.3	—	—	70.0	16.7
—	—	3.3	77.8	18.9
—	13.3	—	43.3	43.3
43.3	—	32.2	—	24.4

Table 2 Cepstral distortion measure

	Whispers	GMM	RBM	DNN	Semi-DNN
mean	8.45	5.37	6.43	5.76	6.06
Standard	4.3	2.61	2.89	2.68	2.62

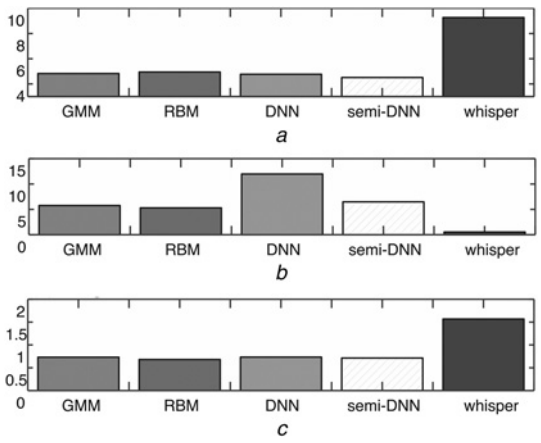

Fig. 4 *Mean objective performance scores obtained from symmetrical IS distance measure, segmental SNR, and LLR*

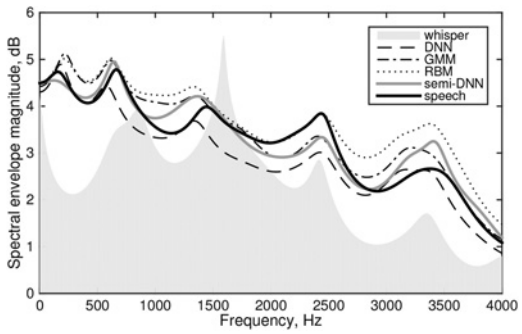

Fig. 5 *Comparison of the spectral envelope shape of each method for a short section of voiced speech*

All models have significantly transformed the whisper spectrum (background shading) to become much more similar to that of the corresponding voiced speech (thick dark line). Assuming that the spectral peaks represent formants, then formant location and amplitude of the reconstructed output closely follows that of the speech. The semi-DNN output is arguably slightly closer than the other methods (apart from the peak at about 2.5 kHz which the RBM method matches more closely).

A different utterance is explored in Fig. 6, with spectrograms of each of the six signals plotted for comparison. The different timescale on the speech spectrum is needed to align it to the whisper and reconstructed speech which are of a different duration. Note the finely detailed formant tracks throughout the recording, and the improved contrast of the semi-DNN formant spectra.

11. Conclusion and future work: This Letter has proposed a new method of constructing a DNN for performing statistical WSC. This has been compared and evaluated against a direct standard DNN implementation, as well as against state-of-the-art GMM and RBM methods, demonstrating excellent performance for both subjective and objective criteria. However, the major benefit of the system is that, for the first time, it enables partially supervised training of a statistical WSC DNN system. This is important for enabling future healthcare implementations. Having already suffered voice-loss, patients are unable to record the extensive and high-quality parallel speech and whisper utterance databases required for DNN, GMM, and RBM training. The proposed semi-DNN system still requires parallel training data, but only needs enough to train a single mapping network, rather than an entire system. The high-dimensional feature coding input and output layers can be trained using different material that is not

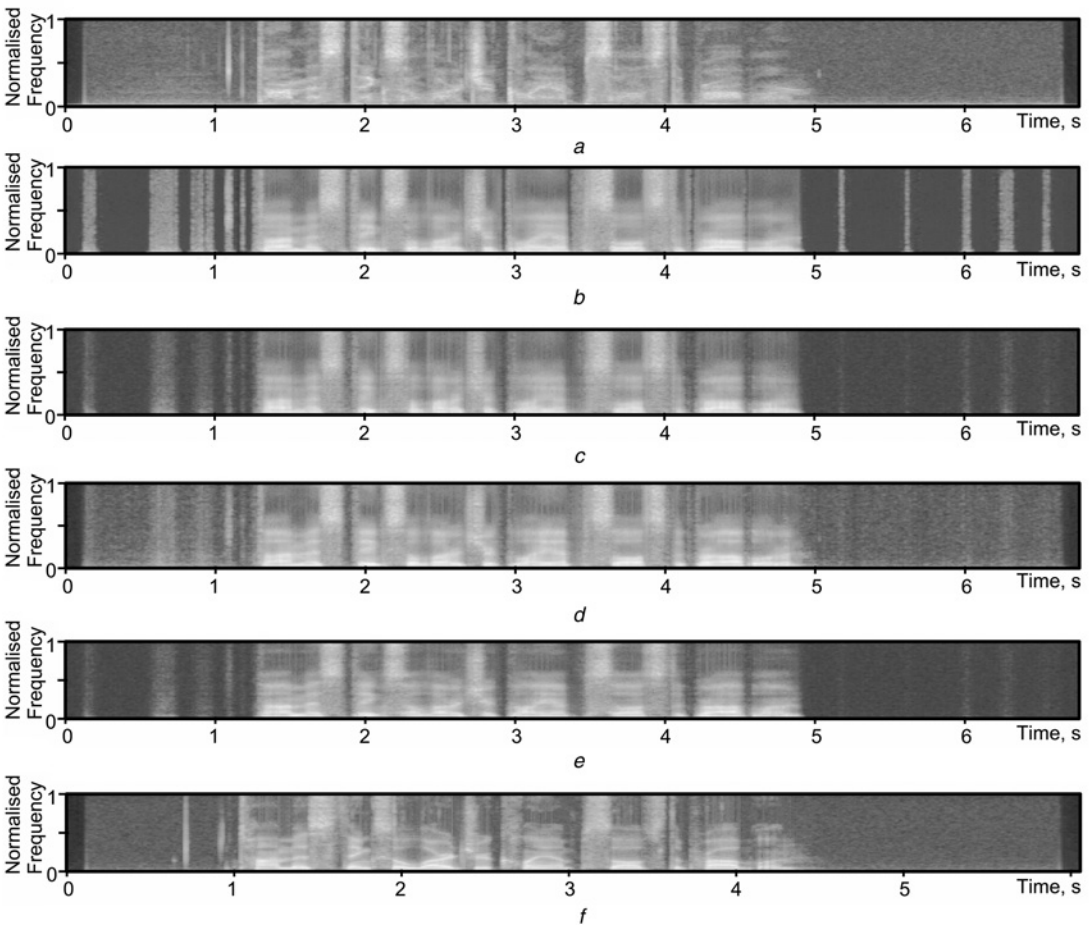

Fig. 6 *Spectrogram plots of*
a Original whispers, speech reconstructed using
b GMM
c DNN
d RBM
e Semi- DNN methods
f Matching speech aligned below

speaker specific. This helps to unlock the potential of high-quality DNN methods for future practical WSC systems.

12. Funding and declaration of interests: Conflict of interest: none declared.

13 References

[1] Mcloughlin I.V., Sharifzadeh H.R., Tan S.L., *ET AL.*: 'Reconstruction of phonated speech from whispers using formant-derived plausible pitch modulation', *ACM Trans. Accessible Comput. (TACCESS)*, 2015, **6**, (4), p. 12

[2] Sharifzadeh H.R., McLoughlin I.V., Ahmadi F.: 'Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec', *IEEE Trans. Biomed. Eng.*, 2010, **57**, pp. 2448–2458

[3] Morris R.W., Clements M.A.: 'Reconstruction of speech from whispers', *Med. Eng. Phys.*, 2002, **24**, (7), pp. 515–520

[4] McLoughlin I.V., Li J., Song Y.: 'Reconstruction of continuous voiced speech from whispers'. Proc. Interspeech, August 2013, pp. 1022–1026

[5] Toda T., Nakagiri M., Shikano K.: 'Statistical voice conversion techniques for body-conducted unvoiced speech enhancement', *IEEE Trans. Audio Speech Lang. Process.*, 2012, **20**, (9), pp. 2505–2517

[6] Li J.-J., McLoughlin I.V., Dai L.-R., *ET AL.*: 'Whisper-to-speech conversion using restricted Boltzmann machine arrays', *Electron. Lett.*, 2014, **50**, (24), pp. 1781–1782

[7] Tajiri Y., Tanaka K., Toda T., *ET AL.*: 'Non-audible murmur enhancement based on statistical conversion using air- and body-conductive microphones in noisy environments'. 16th Annual Conf. of the Int. Speech Communication Association, 2015

[8] Kawahara H., Morise M., Takahashi T., *ET AL.*: 'TANDEM-STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, *f*0, and aperiodicity estimation'. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2008 ICASSP 2008, 2008, pp. 3933–3936

[9] Hinton G., Deng L., Yu D., *ET AL.*: 'Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups', *IEEE Signal Process. Mag.*, 2012, **29**, (6), pp. 82–97

[10] Jiang B., Song Y., Wei S., *ET AL.*: 'Deep bottleneck features for spoken language identification', *PLoS ONE*, 2014, **9**, (7), p. e100795

[11] McLoughlin I., Zhang H.-M., Xie Z.-P., *ET AL.*: 'Robust sound event classification using deep neural networks', *IEEE Trans. Audio Speech Lang. Process.*, 2015, **23**, pp. 540–552

[12] Chen L.-H., Ling Z.-H., Liu L.-J., *ET AL.*: 'Voice conversion using deep neural networks with layer-wise generative training', *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2014, **22**, (12), pp. 1859–1872

[13] Xu Y., Du J., Dai L.-R., *ET AL.*: 'A regression approach to speech enhancement based on deep neural networks', *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2015, **23**, (1), pp. 7–19

[14] Toda T., Black A.W., Tokuda K.: 'Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory', *IEEE Trans. Audio Speech Lang. Process.*, 2007, **15**, (8), pp. 2222–2235

[15] McLoughlin I.V.: 'Speech and audio processing: a MATLAB-based approach' (Cambridge University Press, 2016)