Speech Communication 98 (2018) 42–50

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Updating the Silent Speech Challenge benchmark with deep learning

Ji Yan[a], Liu Licheng[a], Wang Hongcui[b], Liu Zhilei[a], Niu Zhibin[a], Denby Bruce[a,*]

[a] Tianjin University, 135 Yaguan Road, Jinnan District, Tianjin, 300350, China
[b] Zhejiang W.E. University, Hangzhou, China

ARTICLE INFO

Keywords:
Silent speech interface
Multimodal speech recognition
Deep learning
Language model

ABSTRACT

The term "Silent Speech Interface" was introduced almost a decade ago to describe speech communication systems using only non-acoustic sensors, such as electromyography, ultrasound tongue imaging, or electromagnetic articulography. Although the use of specialized sensors in speech processing is challenging, silent speech research remains an active field that can often profit from new developments in traditional acoustic speech processing – for example recent advances in Deep Learning. After an overview of Silent Speech Interfaces and their special challenges, the article presents new results in which a 2010 benchmark study, called the Silent Speech Challenge, is updated with a Deep Learning strategy, using the same input features and decoding strategy as in the original Challenge article. A Word Error Rate of 6.4% is obtained with the new method, compared to the published benchmark value of 17.4%. Additional results comparing new auto-encoder-based features with the original features at reduced dimensionality, as well as decoding scenarios on two different language models, are also presented. The Silent Speech Challenge archive has furthermore been updated to contain both the original and the new auto-encoder features, in addition to the original raw data.

1. Introduction

1.1. Silent Speech Interfaces and challenges

A Silent Speech Interface (Denby et al., 2010), or SSI, is defined as a device enabling speech processing in the absence of an exploitable audio signal – for example, speech recognition obtained exclusively from video images of the mouth, or from electromagnetic articulography sensors (EMA) glued to the tongue. Classic applications targeted by SSIs include:

(1) Voice-replacement for persons who have lost the ability to vocalize through illness or an accident, yet who retain the ability to articulate (see, for example, Fagan et al., 2008 and Denby et al., 2011a);
(2) Speech communication in environments where silence is either necessary or desired: responding to cellphone in meetings or public places without disturbing others; avoiding interference in call centers, conferences and classrooms; private communications by police, military, or business personnel (see, for example, Yuksel et al., 2011).

The SSI concept was first identified as an outgrowth of speech production research, in tandem with the proliferation of the use of cellular telephones, in 2010 in a special issue of Speech Communication (Denby et al., 2010), where SSIs based on seven different non-acoustic sensor types were presented:

(1) MHz range medical ultrasound (US) + video imaging of tongue and lips (Hueber et al., 2010).
(2) Surface electromyography, sEMG, sensors applied to the face and neck (Schultz and Wand, 2010).
(3) Electromagnetic articulography EMA sensors attached to tongue, lips, jaw (Fagan et al., 2008).
(4) Vibration sensors placed on the head and neck (Patil and Hansen, 2010).
(5) Non-audible murmur microphones, NAM, placed on the neck (Tran et al., 2010).
(6) Electro-encephalography, EEG, electrodes (DaSalla et al., 2009).
(7) Cortical implants for a "thought-driven" SSI (Brumberg et al., 2010).

An overview of the SSI concept appears in Fig. 1.

As a non-acoustic technology, SSIs initially stood somewhat apart from the main body of speech processing, where the standard techniques are intrinsically associated with an audio signal. Nevertheless, the novelty of the SSI concept and their exciting range of applications – perhaps aided by an accrued interest in multi-modal speech processing

* Corresponding author.
E-mail address: denby@ieee.org (B. Denby).

https://doi.org/10.1016/j.specom.2018.02.002
Received 9 March 2017; Received in revised form 19 January 2018; Accepted 20 February 2018
Available online 21 February 2018

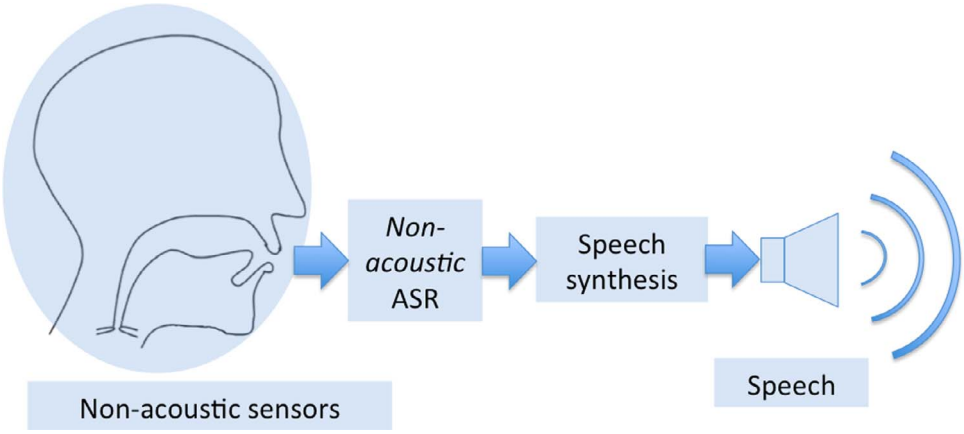


Fig. 1. Overview of an SSI, showing non-acoustic sensors and non-acoustic automatic speech recognition, ASR, which can be followed by speech synthesis, or retained as a phonetic, text, or other digital representation, depending on the application.

– are gradually allowing SSI technology to join the speech processing main stream. Activity in SSI research has remained strong since the publication of Denby et al. (2010), which received the ISCA/Eurasip Best Paper Award in 2015. A recent survey of the literature reveals dozens of publications on SSI systems, using not only on the original seven non-acoustic technologies mentioned above, but also two additional ones, namely, low frequency air-borne ultrasound; and micropower radar (Denby et al., 2010; Fagan et al., 2008; Denby et al., 2011a,b; Yuksel et al., 2011; Hueber et al., 2010, 2007, 2008a,b, 2009, 2011, 2012; Schultz and Wand, 2010; Patil and Hansen, 2010; Tran et al., 2010; DaSalla et al., 2009; Brumberg et al., 2010; Hueber, 2009; Hirahara et al., 2010; Toth et al., 2009; Florescu et al., 2010; Lopez-Larraz et al., 2010; Wand and Schultz, 2011; Freitas et al., 2011, 2013, 2014a,b; Jorgensen et al., 2012; Wang et al., 2012; García et al., 2012; Barbulescu et al., 2013; Hofe et al., 2013; Wand et al., 2013, Gonzalez et al., 2014; Bocquelet et al., 2014; Wang et al., 2014; Salama et al., 2014; Sahni et al., 2014; Balwani et al., 2014; Matsumoto, 2014; Freitas et al., 2015; Al Safi and Alhafadhi, 2015; Bocquelet et al., 2015; Wang et al., 2015; Hahm and Wang, 2015; Jaumard-Hakoun et al., 2015; Xu et al., 2015; Yamaguchi et al., 2015; Hueber and Bailly, 2016; Li, 2016; Cheah et al., 2016; Patil et al., 2016; Yamazaki, 2016).

Despite this activity, SSIs today remain for the most part specialized laboratory instruments. The performance of any automatic speech recognition (ASR) system is most often characterized by a Word Error Rate, or WER, expressed as a percentage of the total number of words appearing in a corpus. To date, no SSI ASR system has been able to achieve WER parity with state-of-the-art acoustic ASR. While the ultrasound SSI system described below in Section 1.2 achieved a 17% WER on a single speaker task, for instance, a state-of-the art commercial recognition system would be expected to obtain a WER of a few percent on such a task. Indeed, a number of practical issues make SSI ASR systems considerably more involved to implement than their acoustic counterparts:

1. **Sensor handling**. While in acoustic ASR this may amount to no more than routine microphone protocol, SSIs' non-acoustic sensors are often rather specialized (and expensive), and require physical contact with, or at a minimum careful placement with respect to, the speech biosignal-producing organs. This introduces problems of invasiveness; non-portability; and non-repeatability of sensor placement, bringing added complexity to SSI experiments.
2. **Interference**. An SSI should in principle be silent, but certain SSI modalities – vibration sensors, radar, and low frequency air-borne ultrasound, for example – are actually associated with signals that can propagate beyond the area of utilization of the SSI. The possibility of interference or interception may limit the adoption of these modalities outside the laboratory.
3. **Feature extraction**. While easily calculated Mel Frequency Cepstral Coefficients, MFCC, have been the acoustic ASR features of choice for decades, feature selection for the specialized sensors of SSIs remains an open question, particularly since many SSI modalities – ultrasound imaging, or EEG, for example – are of much higher intrinsic dimensionality than a simple acoustic signal. Furthermore, while the identification of stable phonetic signatures in acoustic data is today a mature field, the existence of salient landmarks in speech biosignals – arising from imaging modalities or electromyography, for example – is less evident.

The medical US modality, operating in the MHz frequency range, nonetheless fares rather well with respect to the concerns raised above, for a number of reasons. It does not, for example, propagate outside the body. It is also a well established (Stone et al., 1983) and documented (Stone, 2005) technique in speech production and speech pathology research, whose first use in the context of SSIs was discussed in Denby and Stone (2004). US is a also relatively non-invasive modality, requiring only a transducer placed under the speaker's chin, coupled with a small video camera in front of the mouth to capture lip movement. These sensors can be easily accommodated in a lightweight acquisition helmet, thus minimizing sensor placement issues. US tongue imaging, with added lip video, is thus in many ways an attractive modality for building a practical SSI, and it will serve as the basis of the present article.

1.2. The Silent Speech Challenge benchmark

In 2010, an US + lip video SSI trained on the well-known TIMIT corpus achieved, with the aid of a language model (LM), a single speaker WER of 17.4% (84.2% "correct" word rate) on an independent test corpus (Cai et al., 2011), representing a promising early SSI result on a benchmark continuous speech recognition task. Subsequently, the raw image data of Cai et al. (2011), that is, the original tongue ultrasound and lip videos, were made available online as the so-called Silent Speech Challenge, or SSC archive (Denby et al., 2013). The purpose of the archive is to provide a stable data set to which newly developed feature extraction and speech recognition techniques can be applied. The SSC data will serve as the basis of all the experiments reported in this article.

Although a 17.4% WER for an SSI trained on a mono-speaker TIMIT corpus is "promising", it must be remembered that standard acoustic ASR can obtain similar or superior scores after training on the full **multi-speaker** acoustic TIMIT corpus, a much more challenging task. Further advances are thus still necessary in order to truly put Silent

Speech Recognition, SSR, on a par with acoustic ASR.

1.3. Deep learning for Silent Speech Interfaces

In the past several years, improvements in acoustic speech recognition using Deep Neural Network–Hidden Markov Model (DNN–HMM) systems, rather than the traditional Gaussian Mixture Model–HMM (GMM–HMM), have become common. In this approach, a deep learning strategy is used to improve estimation of the emission probabilities of the HMM used for speech decoding. It is natural to ask to what extent a DNN–HMM approach can improve SSR performance as well. Despite the SSI implementation challenges outlined earlier, applications of deep learning techniques to SSR have indeed begun to appear. In Wand and Schultz (2014), for example, tests are reported of phonetic feature discrimination for an EMG-based SSI, without a LM, on a small, experiment-specific speech corpus. In Hahm et al. (2015), deep learning on an EMA based SSI is explored, giving SSR *phone* error rates, PER, of 36%, when the Mocha-TIMIT corpus is used for training, testing, and the development of a specific bigram LM. In Liu et al. (2016), a DNN–HMM is applied to the SSC benchmark data, albeit with a 38% WER, in a study comparing the efficacy of different feature extraction methods.

"Feature-free" approaches to pattern recognition in speech, signal and image processing, based on convolutional neural networks, CNN, have also proven very effective in recent years (Bottou et al., 1989; Le Cun et al., 1998; Krizhevsky et al., 2012; Abdel-Hamid et al., 2014). The CNN is a multilayer neural network consisting of multiple sub-networks with shared weights and overlapping receptive fields, alternated with "pooling" layers that reduce dimensionality by retaining only a subset of the afferent inputs. The use of shared weights across different instances of the identical sub-networks greatly reduces the number of weights to be learned, thus allowing the training of a CNN to remain relatively tractable. CNNs are thought to be able to learn a hierarchy of features, of progressively higher order as information pass from the input to the output of the network.

CNN have, naturally, also begun to make their entry into the field of SSI. In Ephrat and Peleg (2017), which is actually a lip-reading application, a CNN is trained to transform video frames from a large video database directly into synthesized un-vocalized speech, using the video sound track to create source-filter type training labels. In Tatulli and Hueber (2017), a CNN is trained to recognize phonetic targets in US tongue and video lip images in a 488-sentence single speaker database, using a phonetically labeled sound track as ground truth, for a speech recognition task with an HMM-GMM. In Xu et al. (2017), CNN are used to recognize tongue gestural targets in US tongue images for an isolated phoneme + nonsense word recognition task. In the latter reference, extensive use is made of data augmentation (Krizhevsky et al., 2012) to increase the size of the training set, often a concern in using CNN, which require very large training sets to be effective, due to the large number of weights that must be learned. Conceivably, the CNN technique could be tested on the raw images of the SSC archive, As the archive contains no sound track, however, pre-training of the CNN, as in Ephrat and Peleg (2017) and Tatulli and Hueber (2017), will not be feasible: for the Challenge data, the CNN training will have to take place conjointly with that of the HMM probabilities. A study of this possibility will appear in an upcoming article.

The present article reports on the first application of the DNN–HMM approach to the SSC recognition benchmark using the same input features and decoding strategy as those reported in Cai et al. (2011), thus allowing a direct comparison of performances. The SSR results obtained here are significantly improved compared to the archive, giving, in the best scenario, a 6.4% WER (94.1% "correct" word recognition rate), or a nearly threefold improvement over the benchmark value. In contrast to Wand and Schultz (2014) and Hahm et al. (2015), furthermore, the LM used in Cai et al. (2011), also employed here, was developed on a completely independent speech corpus. In addition, results with a second, less task-specific LM are included in the present article. Finally, tests of reduced dimensionality feature vectors, as well as completely new input features created from raw SSC archive data, are also reported here. All new features have been added to the SSC archive for future use by other researchers.

In the remainder of the article, the details of the SSC data acquisition system and a description of the available archive data are first summarized, in Section 2. Section 3 then describes the feature extraction strategy developed for the present study; while full details of the DNN–HMM based speech recognition procedure appear in Section 4. The results are summarized in Section 5, and some conclusions and perspectives for future work outlined in the final section.

2. SSC data acquisition and archive resources

The SSC data acquisition system consisted of an acquisition helmet holding a 128 element, 4–8 MHz US probe for tongue imaging, and a black and white, infrared-illuminated video camera to capture the lips. The 320 × 240 pixel tongue images and 640 × 480 pixel lip images created by the system were acquired in a synchronized manner at 60 frames per second (fps) using the Ultraspeech multisensory acquisition system (Hueber et al., 2008c).

The SSC training corpus consists of US and lip video data from a single native English speaker pronouncing the 2342 utterances (47 lists of 50 sentences) of the TIMIT corpus, in the non-verbalized punctuation manner. The speech was recorded silently, i.e., without any vocalization; there is therefore no audio track. The test set is comprised of one hundred short sentences selected from the WSJ0 5000-word corpus (Garofalo et al., 1993) read by the same speaker. The data are available at the web address indicated in Denby et al. (2013). The archive initially contained only the raw ultrasound and lip images of the training and test sets; the original features used, as well as the reduced-length feature vectors and new features created for the present article (see Section 3), have now been appended to it. Speech recognition for the Challenge data was carried out in a standard GMM–HMM scheme and made use of a LM, which is also included in the archive. Further details appear in Section 4.

3. Feature extraction

3.1. Introduction

As mentioned earlier, speech recognition from non-acoustic sensor data faces the problem of discovering an effective feature recognition strategy, and US + lip video SSIs, although attractive in many ways, share this drawback. Being based on images, their intrinsic input dimensionality may be of the order of 1 million pixels. Some means of dimension-reducing feature extraction is thus critical. (The following discussion is centered on tongue features. Lip features, which are much easier to handle, will for overall coherence be extracted in the same way as tongue features.)

3.2. Difficulty of using a contour extraction approach

Tongue contour extraction is a tempting choice for reducing dimensionality that retains visual interpretability of the features. In ultrasound imaging of the tongue, the air-tissue boundary at the upper surface of the tongue produces a bright, continuous contour, referred to in a side-looking scan as the sagittal contour. Image processing tools for automatically extracting and characterizing this contour make ultrasound imaging a powerful tool for the study of speech production (Stone et al., 1983; Stone, 2005). Unfortunately, despite extensive literature on techniques for extracting tongue contours from ultrasound data (see Li et al., 2005; Tang and Hamarneh, 2010; Xu et al., 2016, and references therein), tongue contour tracking remains an extremely challenging task. The high level of speckle noise in ultrasound images

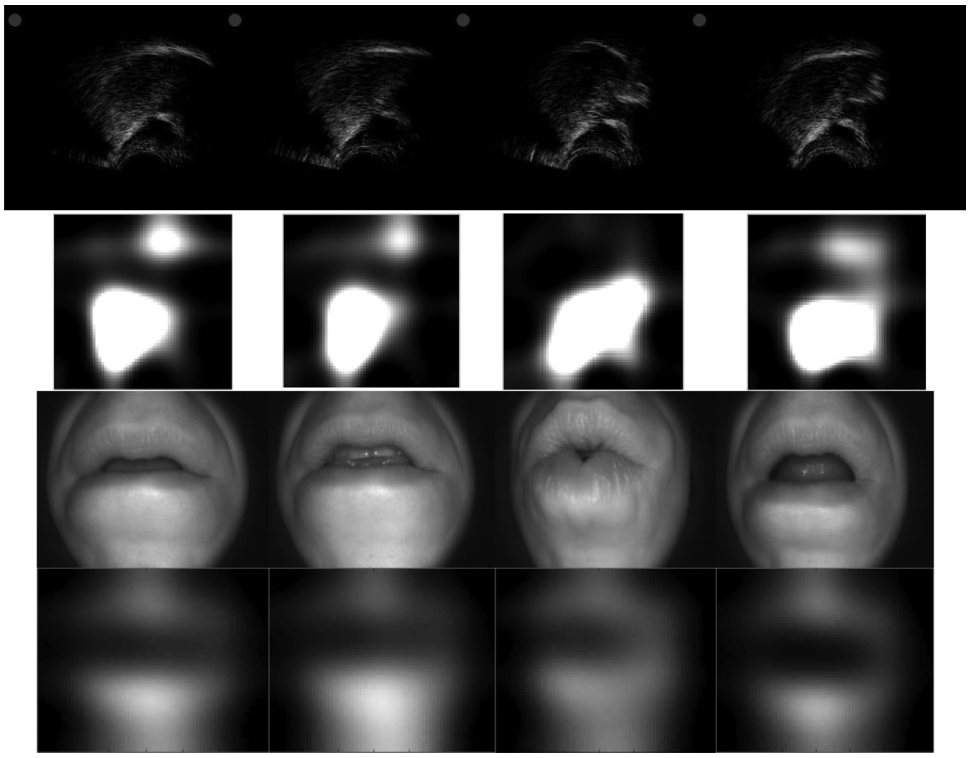


Fig. 2. Original lip (top row) and tongue (third row) images compared to their reconstructions (second and fourth rows, respectively) using 30 DCT coefficients.

(multiplicative noise arising from the coherent nature of the ultrasound wave), coupled with variations in acoustic contact between the transducer and the speaker's skin; blocking of the ultrasound wave by the hyoid bone and jaw; poor reflectivity of muscle fibers in certain orientations of the tongue; and the lack of a complete echogenic tissue path to all parts of the tongue surface, in particular the tongue tip, often result in sagittal contours that are incomplete, contain significant artifacts, or are even totally absent. While even imperfect automatically-extracted contours remove the tedium of hand-scanning and are valuable for qualitative studies, it is difficult to integrate such information in a coherent way into labeled training datasets intended for machine learning tasks such as speech recognition. As a consequence, US-based SSIs have tended to use projective feature extraction techniques rather than contour finding. In work performed thus far, Principal Component Analysis, PCA, and the Discrete Cosine Transform, DCT, have been the methods of choice.

3.3. PCA and DCT approaches

In Hueber et al. (2007, 2008a), PCA was used on ultrasound-based SSIs in an "Eigentongues" approach, wherein each ultrasound image is represented as a linear combination of a set of orthogonal Eigen-images determined on a training set of representative images. The first 30 or so Eigentongues were found sufficient to represent the discriminative power contained in the ultrasound images (Hueber, 2009).

The DCT, widely used for lossy image compression, is based on the notion that most of the information in an image is concentrated in the lower spatial frequencies (Rao and Yip, 1990). We note that the DCT, as a direct multiplicative transform related to the Fourier transform, does not make use of a training set. The technique for calculating the DCT coefficients will be presented in the next section. In Cai et al. (2011), the article on which the SSC archive is based, it was found that DCT features provided substantially better recognition scores, as well as faster execution times, than the Eigentongue approach. **This result leads to the important consequence that the SSC benchmark refers to recognition scores obtained using DCT features** (we note in addition that the original Eigentongue features of (Cai et al., 2011) are no longer available). Consequently, a quantitative comparison of a DNN–HMM approach to the GMM–HMM analysis used in the original benchmark – which is the major impetus of this article – must make use of the identical DCT features in its baseline result.

The SSC archive DCT features were constructed in the following way. First, fixed Regions of Interest (ROI) of tongue and lip images were resized to 64 by 64 pixels. This resizing is necessary in order to keep the number of DCT coefficients tractable. For an image matrix A of size $N*N$, the two-dimension DCT is then computed as:

$$D_{ij} = a_i a_j \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} A_{mn} \cos \frac{\pi(2m+1)i}{2N} \cos \frac{\pi(2n+1)j}{2N} \tag{1}$$

$$0 \ll \mathrm{i, j} \leq \mathrm{N} - 1$$

where

$$a_i,\, a_j = \begin{cases} \dfrac{1}{\sqrt{N}}, & i, j = 0 \\ \sqrt{\dfrac{2}{N}}, & 1 \ll i, j \ll N - 1 \end{cases}$$

Dimensionality reduction is achieved by retaining only the K lowest frequency DCT components. In Cai et al. (2011), a feature size of $K = 30$ was selected, based on performance comparisons. In acoustic speech recognition, it is usual to concatenate the first derivative, or Δ, of the MFCC feature vector to the vector itself. This procedure was also carried out for the DCT features of the archive, thus creating a 120-component feature vector for each tongue + lip frame.

3.4. New features created with Deep Auto Encoder

Although DCT features have provided promising recognition results for US + lip video SSIs, it has been necessary to make certain compromises in extracting them, notably: 1) resizing the original images before calculating them; and 2) retaining only a small, fixed number of DCT coefficients. While computational tractability issues prevent us, at

present, from removing the first restriction, the presence of the raw tongue and lip data in the SSC archive allows us to consider taking a closer look at the second one.

It is first of all interesting to examine the appearance of tongue and lip images reconstructed using 30 DCT coefficients. An example result on 4 frames is given in Fig. 2.

Although the lip reconstructions are sufficiently clear to distinguish an overall degree of mouth opening, an acoustically pertinent quantity, the visual fidelity of the tongue images is rather poor. The information in the tongue images necessary for distinguishing different acoustic configurations is, evidently, coded by the DCT in a way that does not retain a high level of visual fidelity. It is tantalizing to ask, however, whether one might do better by creating, from the original images present in archive, a new feature representation that reduces dimensionality while explicitly preserving visual fidelity, rather than relying on a somewhat arbitrarily placed cut in spatial frequency space, as was done for the DCT. A Deep Auto Encoder, or DAE, was used to explore this possibility.

A Deep Auto-Encoder is a neural network used for reducing the dimensionality and learning a representation of input data (Hinton, 2010). It contains an "encoder" and a "decoder" symmetrically arranged about a "code" layer, as shown in Fig. 3. The action of the encoder can be defined as:

$$z = f(Wx + b) \tag{2}$$

where f is an activation function, such as sigmoid function, W a weight matrix, and b a bias. The decoder output is defined as:

$$x' = f(W'z + b') \tag{3}$$

where x' is of the same dimension as x. The weight matrix W' is equal to W^T. An autoencoder is trained by minimizing the image reconstruction error, computed as:

$$L(x,\ x') = -\sum_{k=1}^{d} [f_1(x_k,\ x_k^{'}) + f_2(x_k,\ x_k^{'})] \tag{4}$$

where

$$f_1(x,\ x') = x \log x'$$

$$f_2(x,\ x') = (1 - x)\log(1 - x').$$

When training is complete, the code layer may be regarded as a compressed representation of the input, and is suitable for use as a feature vector. Details of the DAE training procedure can be found in Hinton and Salakhutdinov (2006).

To obtain the new features, ROIs were selected and resized, once again for computational tractability purposes, via bi-cubic interpolation, to 50 × 70 (lip) and 50 × 60 (tongue) pixel arrays, which form the inputs to the DAE. We note that the original ROIs from the Challenge article were not available. The choice made for the DAE gives nearly the same resolution as the ROIs of the Challenge article while slightly reducing dimensionality by preserving the aspect ratio of the ROI during resize. A direct comparison with DCT remains meaningful, since 1) neither set of ROIs lost any significant data; and 2) the retained low frequency DCTs are quite insensitive to the original image resolution. After tests with various architectures, a J-1000-500-250-K network was chosen, with J the number of inputs (3500 for lips and 3000 for the tongue); K the desired dimensionality of the created feature vector; and the intermediate figures the number of neurons in each layer. Features were calculated for K = 30, 20, 10, and 5. The encoder and symmetric decoder networks were trained on 12 lists of images (12 x 50 = 600 images) selected at random from the SSC TIMIT training corpus.

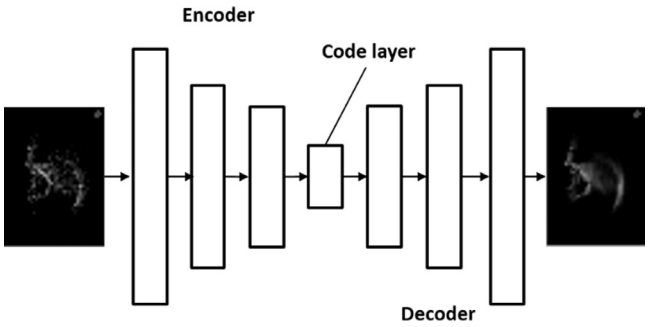


Fig. 3. Architecture of DAE.

Reconstructed images (bottom row of each panel) for tongue and lips are compared with the original images (top row of each panel) in Fig. 4, where (a) and (b) show the results using 30 and 5 DAE features respectively. The figure shows that remarkable visual fidelity can be obtained using only 5 DAE features. This is in contrast to images reconstructed using DCT features shown previously, which are barely recognizable even for the 30 dimensional features. Although one may ask to what extent the DAE solution is similar to PCA, the K = 5 case, with, as will be seen later, the SSR results it allows to obtain, is nonetheless intriguing.

4. DNN–HMM speech recognition

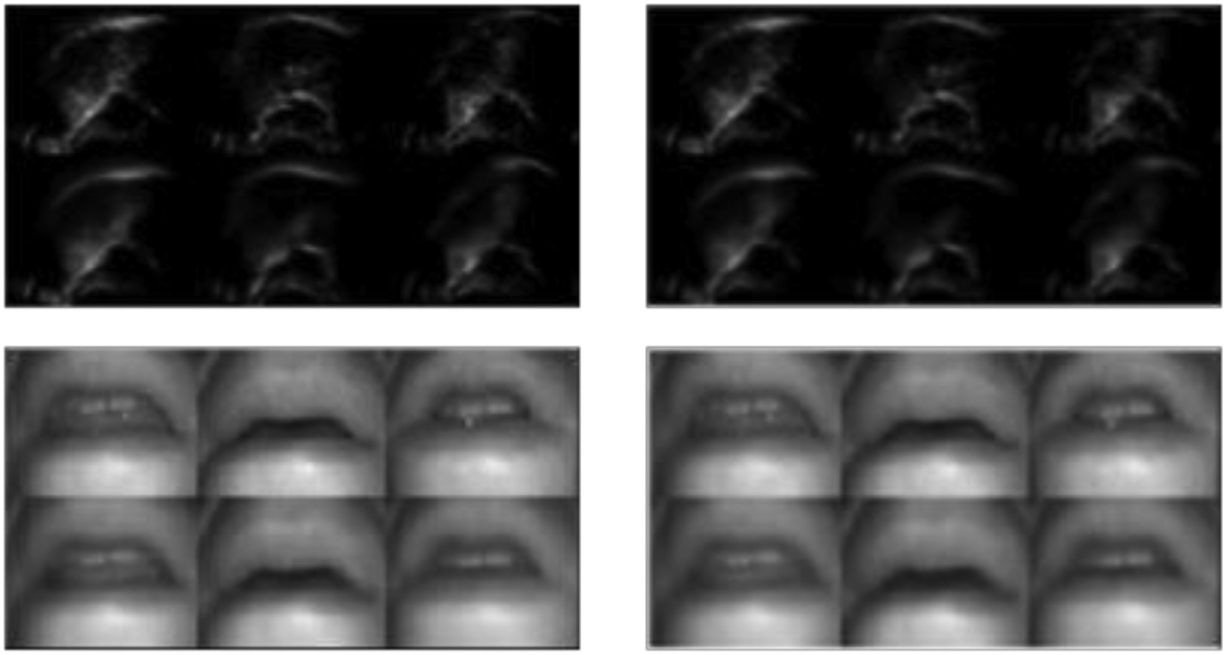
4.1. System overview

The Kaldi open-source Deep Learning toolkit (Povey et al., 2011; The Kaldi Toolkit) was used to build the SSR system, whose overall architecture is illustrated in Fig. 5 and described in more detail in the following paragraphs. Features were normalized to have zero mean and unit variance before use.

We note that in the SSC benchmark, HTK was used to perform the speech recognition, using a standard GMM–HMM architecture. In order to ensure as meaningful a comparison as possible with the benchmark result, without actually *re-doing* it with HTK, the recognition with Kaldi was performed first using a GMM–HMM, and then a DNN–HMM. The procedures used for the non-acoustic ASR, were adapted from standard recipes in acoustic speech recognition and Deep Learning (Dahl et al., 2015, 2011; Hinton et al., 2012; Popović et al., 2015; Rath et al., 2013), in particular "Karel's Implementation" (Rath et al., 2013), modified to accept non-acoustic input features. They are described below.

In the Kaldi GMM–HMM "acoustic" model training stage (the name "acoustic model" is retained even though the input feature data used here are non-acoustic), a monophone model was first trained using combined tongue and lip feature vectors, of type DCT or DAE, of dimension K = 30, 20, 10, and 5. The monophone pass used flat start and 40 iterations, and had a total of 1700 Gaussians. Subsequently, two triphone passes were performed, each of 35 iterations. In the first triphone model, called triphone1, the Δ (i.e., time derivative) features (DCT or DAE, for lips plus tongue) were appended to the original features (Fig. 5), with the alignment from the monophone providing the training labels. This pass comprises 1800 regression tree leaves and 9000 Gaussians. Next, the triphone2b model was created, comprising 3000 regression tree leaves and 25,000 Gaussians. In this step, the feature vector is again modified, this time by applying Linear Discriminant Analysis (LDA) and a Maximum Likelihood Linear Transformation (MLLT) (Fig. 5), so as to replace the Δ features of triphone1 with a new vector of dimension 40. The training of triphone2b was based upon the alignment of the previous triphone2 pass. Thus, the monophone, triphone1 and triphone2b acoustic models were trained consecutively, each time using the previous model for alignment, with WER performance improving on each subsequent pass. At this point, the alignment of triphone2b can be used to train the DNN–HMM, to try to improve the WER even further.

The role of the DNN–HMM, implemented here as a Deep Belief Network (DBN), is to output improved estimates of the HMM observation probabilities for all tied-triphone states on individual frames. This is obtained by training the DNN via a cross-entropy criterion to



(a) $K = 30$ (b) $K = 5$

Fig. 4. Original (top row of each panel) and reconstructed (bottom row of each panel) images of tongue and lips using two dimensionalities of DAE features.

classify frames into triphone states, using the triphone2b alignment as training labels, then transforming the network outputs into probabilities with a softmax output layer. The new HMM observation probabilities replace those obtained in previous passes (Fig. 5), and final DNN–HMM decoding performed in the classical way (the HMM transition probabilities are those from the previous GMM–HMM phase). The DBN implemented is illustrated in Fig. 6 (Dahl et al., 2015, 2011; Hinton et al., 2012; Popović et al., 2015; Rath et al., 2013), using the K-dimensional DCT or DAE features as input, over an 11 frame (5 before and 5 after the current frame) context window. Restricted Boltzmann Machines (RBM) are cascaded by means of the weight vector W in the figure. The DNN training operates in two phases. During the pre-training phase, the RBMs are trained in an unsupervised way using the contrastive divergence (CD) algorithm. The six, 1024-unit hidden layers of the RBMs are made up of Bernoulli-Bernoulli units (learning rate 0.4) except for the first, which is Gaussian-Bernoulli (learning rate 0.01). The weights learned in pre-training phase are then used to initialize the DNN model. In the second DNN training phase, 90% of the training set is used for training, optimizing per-frame cross-entropy, and the remaining 10% of the training data for evaluation, cf. Popović et al. (2015). Four hidden layers of 1024 sigmoid units each were used in this phase, with the DNN softmax output layer estimating observation probabilities for the 2403 pruned tied HMM states. The DNN architecture was implemented on a CUDA GPU machine. System parameters used, including the DNN parameters, total numbers of Gaussians, tied states (regression tree leaves), search space (beam), and "acoustic"/LM weight (acwt) parameters, are summarized in Table 1.

4.2. Language model and lexicon issues

The WSJ0 5k NVP bigram LM (lm_wsj_5k_nvp_2gram, Garofalo et al., 1993), used in the decoding stage of the SSC benchmark and derived from a fixed 5000-word subset of the Wall Street Journal (WSJ) text corpus, was also adopted in these tests. Obtaining realistic WER scores on small corpora, however, can be problematic. Using a closed vocabulary, as is the case here, tends towards underestimation of attainable WER. On the other hand, an unbiased lexicon derived exclusively from a small training set might not contain all the words present in the test set, thus leading to an overly pessimistic WER estimate. To help address these issues, a second estimate of the achievable WER on these data was also made using another, less task-specific CSR 5k NVP bigram LM (lm_csr_5k_nvp_2gram, Garofalo et al., 1993). This LM contains newswire text from WSJ, the San José Meteor, and the Associated Press, along with some spontaneous dictation by journalists of hypothetical news articles. Results on both LM appear in the next section.

5. Results and analysis

Table 2 shows a comparison of the Kaldi DNN–HMM results, on the

Fig. 5. Overall SSR training procedure. The upper branch shows the GMM/HMM monophone and two types of triphone training. In the lower branch, the DNN, which uses the same features as the GMM/HMM model, after setting initial parameters, undergoes first an unsupervised pretraining, then a supervised training; it has as its output estimated observation probabilities that are fed (vertical arrow) back into the second triphone model for the final decoding step. See text for a more complete description.

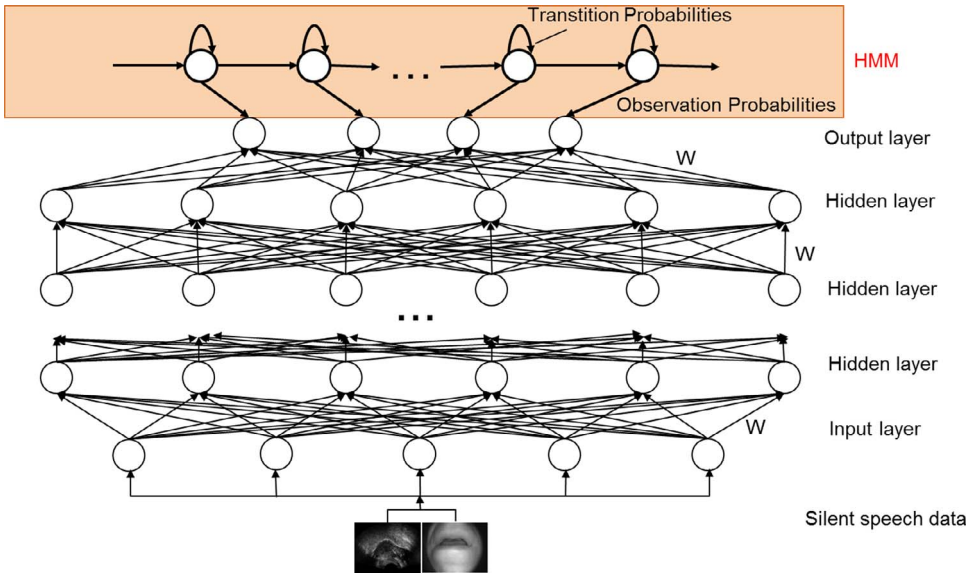


Fig. 6. DNN structure used for SSR.

Table 1
SSR system parameters.

Monophone	Tot_Gaussian	1700
Triphone1	Regression tree leaves	1800
	Tot_Gaussian	9000
Triphone2b	Regression tree leaves	3000
	Tot_Gaussian	25,000
DNN pretrain	Number of hidden layers	6
	Units per hidden layer	1024
DNN training	Number of hidden layers	4
	Units per hidden layer	1024
	beam	13.0
	Lattice_beam	8.0
	acwt	0.1

Table 2
Comparison with original HTK result of Patil et al. (2016), using 30-element DCT features.

ASR system	HTK SSC benchmark	Kaldi GMM–HMM	Kaldi DNN–HMM
WER	17.4%	17.4%	**6.45%**

WSJ0 5000-word corpus, to those of the SSC benchmark, using the same 30-dimensional DCT input features and decoding strategy as in Cai et al. (2011). Although a test with HTK itself was not repeated here, the fact that quite similar results were obtained using a GMM–HMM in Kaldi (column 2) provides reassurance that the figures obtained using Kaldi are reasonable. The Table shows that the DNN–HMM strategy has reduced the WER by almost a factor of 3 as compared to the benchmark.

To perform the LM tests proposed in Section 4.2, the procedure was repeated using the alternate LM, as shown in Table 3 for 30-element

Table 3
Comparing results for the 2 different LM, for 30-element feature vectors of both types.

LM		WER (%)	
		lm_csr_5k_nvp_2gram	lm_wsj_5_nvp_2gram
DCT	monophone	45.55	40.47
	Triphone2b	17.40	12.71
	DNN	11.44	**6.45**
DAE	Monophone	58.55	59.92
	Triphone2b	21.70	14.76
	DNN	13.98	**7.72**

Table 4
Recognition results with WSJ language model.

		Feature vector dimension K			
		30	20	10	5
WER (%) for DCT features	Monophone	40.47	37.15	36.36	98.24
	Triphone2b	13.00	14.76	12.32	100
	DNN	**6.45**	**6.35**	**7.43**	99.51
WER (%) for DAE features	Monophone	59.92	44.18	41.15	45.45
	Triphone2b	14.76	14.96	15.54	17.79
	DNN	**7.72**	**7.72**	**8.80**	**10.07**

Table 5
Recognition results with CSR language model.

		Feature vector dimension K			
		30	20	10	5
WER (%) for DCT features	Monophone	45.55	40.86	39.78	98.34
	Triphone2b	17.79	19.16	16.42	100
	DNN	**11.44**	**11.53**	**12.32**	99.80
WER (%) for DAE features	Monophone	58.55	52.00	49.76	54.25
	Triphone2b	21.70	21.41	19.75	22.48
	DNN	**13.98**	**13.10**	**14.37**	**14.86**

feature vectors of both types (DCT and DAE). One notes first of all that the DCT and DAE features give similar performances, barring the monophone case. We will return to this point in the discussion of Tables 4 and 5. Nonetheless, although the WER performance obtained on the less specific LM is somewhat worse, as expected, it is still significantly better than the SSC benchmark, for both types of features.

To further explore different types of input features, DCT and DAE feature vectors of dimension $K = 20$, 10, 5 for each visual modality were also tested. Results are given in Table 4 for the WSJ LM, and in Table 5 for the CSR LM. Overall, higher scores are again obtained with the more task specific LM, as expected. The tables also show that for both LM, similar results are obtained for the two types of features, with the DCT being slightly better, when the dimensionality K of the input vectors is 10 or more. For $K = 5$, however, while the DCT features are no longer salient, the DAE retains most of its effectiveness. A clue to this behavior is the observation that the DCT performs well even when the image reconstructed from it is poorly recognizable (comparing Figs. 2 and 4), suggesting that visual fidelity is not a salient quantity for the

problem at hand. One may hypothesize that the DCT has "already done some of the work" of the DNN by extracting, apparently salient, frequency information. The DAE, in this interpretation, "unnecessarily" preserves a higher quality version of the input image, "ignoring" other possible types of saliency, perhaps even "forcing" the DNN to *derive* salient information from the DAE features. Then, for the lowest dimensionality, $K = 5$, the DNN may still be able to obtain as much salient information from the DAE features as for other values of K, while the DCT has clearly lost crucial information. Thus, although the DAE has not been completely successful at simultaneously optimizing saliency and low dimensionality, the results it furnishes are intriguing, and suggest that it may be possible to do better with a more sophisticated approach.

6. Conclusions and perspectives

A confrontation of the SSC recognition benchmark with DNN–HMM SSR techniques using the Kaldi Deep Learning package has led to an improvement in WER of almost a factor of 3 in the most favorable scenario, thus helping to establish US as a highly attractive SSI modality. Tests performed using both the original WSJ LM and a less task-specific CSR LM give WER values that are on these data, using the significantly improved compared to the benchmark. Before the DNN–HMM tests, Kaldi was also used to test a GMM–HMM architecture, in order to demonstrate compatibility with the methods used in the benchmark. New features derived from the raw benchmark data using a DAE give results only slightly worse than those obtained with the original DCT features, while retaining their effectiveness even at very low dimensionality. Both new and original features have now been appended to the SSC benchmark data.

While these results are promising, SSR still remains somewhat less accurate than acoustic speech recognition, and further work will be necessary. In the future, for the SSC benchmark, it will be interesting to experiment with other feature extraction strategies, for example convolutional neural networks, CNN, which might allow the image-resizing step, where information may be lost, to be skipped. For SSI more generally, it will be interesting to accumulate much larger (if possible multi-speaker) data sets, so that some of the mentioned problems associated with small speech data sets may be avoided.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 61303109 and No. 61503277) and 985 Foundation from China's Ministry of Education (No. 060-09030710001).

References

Abdel-Hamid, O., Mohamed, A.-R., Jiang, H., Deng, L., Penn, G., Yu, D., 2014. Convolutional neural networks for speech recognition. IEEE/ACM Trans Audio Speech Lang. Process. 22 (10).

Al Safi, A., Alhafadhi, L., 2015. Review of EMG-based speech recognition. Int. J. Rev. Electron. Commun. Eng. 3 (June (3)).

Balwani, D., Brijwani, H., Daswani, K., Rastog, S., 2014. Talking without talking. Int. J. Eng. Res. Appl. 4 (April (4)), 55–56.

Barbulescu, A., Hueber, T., Bailly, G., Ronfard, R., 2013. Audio-visual speaker conversion using prosody features. In: Proceedings of International Conference of Audio-visual Speech Processing (AVSP), Annecy, France.

Bocquelet, D., Hueber, T., Girin, L., Badin, P., Yvert, B., 2014. Robust articulatory speech synthesis using deep neural networks for BCI applications. In: Proceedings of Interspeech, Singapore, Malaysia, pp. 2288–2292.

Bocquelet, D., Hueber, T., Girin, L., Savariaux, C., Yvert, B., 2015. Real-time control of a DNN-based articulatory synthesizer for silent speech conversion: a pilot study. In: Proceedings of Interspeech, Dresden, pp. 2405–2409.

Bottou, L., Fogelman Soulié, F., Blanchet, P., Lienard, J.S., 1989. Experiments with time delay networks and dynamic time warping for speaker independent isolated digits recognition. In: Proceedings of Eurospeech, Paris, France. ISCA Archive.

Brumberg, J.S., Nieto-Castanon, A., Kennedy, P.R., Guenther, F.H., 2010. Brain-computer interfaces for speech communication. Speech Commun. 52 (April (4)).

Cai, J., Denby, B., Roussel-Ragot, P., Dreyfus, G., Crevier-Buchmann, L., 2011. Recognition and real time performance of a lightweight ultrasound based silent speech interface employing a language model. In: Proceedings of Interspeech, Florence, Italy, pp. 1005–1008.

C.C. Jorgensen, D.D. Lee, S.T. Agabon, United States National Aeronautics and Space Association, Silent Speech Sub-Audible Speech Recognition Based on Electromyographic Signals, US Patent number 8200486B1; 2012.

Cheah, L.A., Gilbert, J.M., Gonzalez, J.A., Bai, J., Ell, S.R., Fagan, M.J., Moore, R.K., Green, P.D., Rychenko, S.I., January 2016. Integrating User-Centred Design in the Development of a Silent Speech Interface Based on Permanent Magnetic Articulography, Biomedical Engineering Systems and Technologies. Communications in Computer and Information Science, vol. 574. pp. 324–337.

Dahl, G., Yu, D., Deng, L., Acero, A., 2011. Large vocabulary continuous speech recognition with context-dependent DBN-HMMs. In: Proceedings of ICASSP2011, Prague, Czech Republic, pp. 4688–4691.

Dahl, G., Yu, D., Deng, L., Acero, A., 2015. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. 20 (1), 30–42.

DaSalla, C.S., Kambara, H., Sato, M., Koike, Y., 2009. Spatial filtering and single-trial classification of EEG during vowel speech imagery. In: Proceedings of the 3rd International Convention on Rehabilitation Engineering and Assistive Technology (i-CREATe 2009), Singapore.

Denby, B., Cai, J., Hueber, T., Roussel, P., Dreyfus, G., Crevier-Buchman, L., Pillot-Loiseau, C., Chollet, G., Manitsaris, S., Stone, M., 2011b. Towards a practical silent speech interface based on vocal tract imaging. In: 9th International Seminar on Speech Production (ISSP 2011), Montréal, Canada, pp. 89–94.

Denby, B., Cai, J., Roussel, P., Dreyfus, G., Crevier-Buchman, L., Pillot-Loiseau, C., Hueber, T., Chollet, G., 2011a. Tests of an interactive, phrasebook-style post-laryngectomy voice-replacement system. In: Proceedings of ICPhS, Hong Kong.

Denby, B., Hueber, T., Cai, J., Roussel, P., Crevier-Buchman, L., Manitsaris, S., Chollet, G., Stone, M., Pillot, C, 2013. The silent speech challenge archive. online: https://ftp.espci.fr/pub/sigma/ .

Denby, B., Schultz, T., Honda, K., Hueber, T., M.Gilbert, J., Brumberg, J.S., 2010. Silent speech interfaces. Speech Commun. 52 (April (4)), 270–287.

Denby, B., Stone, M., 2004. Speech synthesis from ultrasound images of the tongue. In: Proceedings of ICASSP, Montréal, Canada.

Ephrat, A., Peleg, S., 2017. ID2Speech: speech reconstruction from silent video. In: Proceedings of ICASSP2017, New Orleans, USA.

Fagan, M.J., Ell, S.R., Gilbert, J.M., Sarrazin, E., Chapman, P.M., 2008. Development of a (silent) speech recognition system for patients following laryngectomy. Med. Eng. Phys. 30 (May (4)), 419–425.

Florescu, M., Crevier-Buchman, L., Denby, B., Hueber, T., Colazo-Simon, A., Pillot-Loiseau, C., 2010. Silent vs vocalized articulation for a portable ultrasound-based silent speech interface. In: Proceedings of Interspeech 2010, Makuhari, Chiba, Japan, September 26-30.

Freitas, J., Ferreira, A.J., Teles de Figueiredo, M.A., Teixeira, A., Dias, M.S., 2014a. Enhancing multimodal silent speech interfaces with feature selection. In: Proceedings of Interspeech, Singapore, Malaysia, pp. 1169–1173.

Freitas, J., Teixeira, A., Dias, M.S., 2013. Multimodal silent speech interface based on video, depth, surface electromyography and ultrasonic doppler: data collection and first recognition results. In: Proceedings of SPASR 2013, Lyon, France.

Freitas, J., Teixeira, A., Dias, M.S., 2014b. Multimodal corpora for silent speech interaction. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland.

Freitas, J., Teixeira, A., Dias, M.S., Bastos, C., 2011. Towards a multimodal silent speech interface for European Portuguese. In: Ipsic, Ivo (Ed.), Speech Technologies. Intechopen. cdn.intechopen.com.

Freitas, J., Teixeira, A., Silva, S., Oliveira, C., Dias, M.S., 2015. Detecting nasal vowels in speech interfaces based on surface electromyography. PLoS One 10 (6).

García, A.A., Reyes-García, C.A., Villaseñor-Pineda, L., 2012. Towards a silent speech interface based on unspoken speech. In: Proceedings of Biosignals 2012 (BIOSTEC), Algarve, Portugal. SciTePress, pp. 370–373.

Garofalo, J., Graff, D., Paul, D., Pallett, D, 1993. CSR-I (WSJ0) complete. online: https://catalog.ldc.upenn.edu/LDC93S6A .

Gonzalez, J.A., Cheah, L.A., Bai, J., Ell, S.R., Gilbert, J.M., Moore, R.K., Green, P.D., 2014. Analysis of phonetic similarity in a silent speech interface based on permanent magnetic articulography. In: Proceedings of Interspeech, Singapore, Malaysia.

Hahm, S., Wang, J., 2015. Silent speech recognition from articulatory movements using deep neural network. In: Proceedings of ICPhS2015, Glasgow.

Hahm, S., Wang, J., Friedman, J., 2015. Silent speech recognition from articulatory movements using deep neural network. In: Proc. of the International Congress of Phonetic Sciences, Glasgow, Scotland, UK, pp. 1–5.

Hinton, G., 2010. A practical guide to training restricted Boltzmann machines. Momentum 9 (1), 92ff.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process. Mag. 29 (6), 82–97.

Hinton, G., Salakhutdinov, R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504–507.

Hirahara, T., Otani, M., Shimizu, S., Toda, M., Nakamura, K., Nakajima, Y., Shikano, K., 2010. Silent-speech enhancement system utilizing body-conducted vocal-tract resonance signals. Speech Commun. 52 (4).

Hofe, R., Ell, S.R., Fagan, M.J., Gilbert, J.M., Green, P.D., Moore, R.K., Rybchenko, S.I., 2013. Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. Speech Commun. 55 (1), 22–32.

Hueber, T., December 2009. Reconstitution De La Parole Par Imagerie Ultrasonore Et Vidéo De L'appareil Vocal: Vers Une Communication Parlée Silencieuse. Université Pierre et Marie Curie (in French), Doctoral Thesis.

Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., Stone, M., 2007. Eigentongue feature extraction for an ultrasound-based silent speech interface. In: Proceedings of ICASSP, Honolulu, USA, pp. 1245–1248.
Hueber, T., Badin, P., Bailly, G., Ben-Youssef, A., Elisei, F., Denby, B., Chollet, G., 2011. Statistical mapping between articulatory and acoustic data, application to silent speech interface and visual articulatory feedback. In: Proceedings of Interspeech, Firenze, Italy, pp. 593–596.
Hueber, T., Bailly, G., 2016. Statistical conversion of silent articulation into audible speech using full-covariance HMM. Comput. Speech Lang. 36, 274–293.
Hueber, T., Bailly, G., Denby, B., 2012. Continuous articulatory-to-acoustic mapping using phone-based trajectory HMM for a silent speech interface. In: Proceedings of Interspeech, Portland, USA.
Hueber, T., Benaroya, E-L., Chollet, G., Denby, B., Dreyfus, G., Stone, M., 2010. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. Speech Commun. 52 (April (4)), 288–300.
Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., 2008b. Phone recognition from ultrasound and optical video sequences for a silent speech interface. In: Proceedings of Interspeech, Brisbane, Australia, pp. 2032–2035.
Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., 2009. Visuo-phonetic decoding using multi-stream and context-dependent models for an ultrasound-based silent speech interface. In: Proceedings of Interspeech, Brighton, UK, pp. 640–643.
Hueber, T., Chollet, G., Denby, B., Stone, M., 2008c. Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. In: Proceedings of the International Seminar on Speech Production, Strasbourg, France, pp. 365–369.
Hueber, T., Chollet, G., Denby, B., Stone, M., 2008a. Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. In: Proceedings of International Seminar on Speech Production, Strasbourg, France, pp. 365–369.
Jaumard-Hakoun, A., Xu, K., Roussel-Ragot, P., Dreyfus, G., Stone, M., Denby, B., 2015. Tongue contour extraction from ultrasound images based on deep neural network. In: Proceedings of ICPhS2015, Glasgow.
Krizhevsky, A., Sutskever, I., Hinton, G., 2012. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, vol. 25. MIT Press, Lake Tahoe, USA, pp. 1097–1105.
Le Cun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.
Li, M., Kambhamettu, C., Stone, M., 2005. Automatic contour tracking in ultrasound images. Clin. Linguist. Phonetics 19 (6-7), 545–554.
Li, W., 2016. Silent speech interface design methodology and case study. Chin. J. Electron. 25 (CJE-1), 88–92.
Liu, L., Ji, Y., Wang, H., Denby, B., 2016. Comparison of DCT and autoencoder-based features for DNN-HMM multimodal silent speech recognition. In: Proceedings of ISCSLP, Tianjin, China.
Lopez-Larraz, E., Mozos, O.M., Antelis, J.M., Minguez, J., 2010. Syllable-based speech recognition using EMG. In: Proceedings of the IEEE Engineering in Medicine and Biology Society, pp. 4699–4702.
Matsumoto, M., 2014. Brain computer interface using silent speech for speech assistive device. PhD Thesis. Doctoral Program in Information Science and Engineering, Graduate School of Science and Technology. Niigata University.
Patil, P., Gujarathi, G., Sonawane, G., 2016. Different approaches for artifact removal in electromyography based silent speech interface. Int. J. Sci. Eng. Technol. Res. 5 (January (1)).
Patil, S.A., Hansen, J.H.L., 2010. A competitive alternative for speaker assessment: physiological microphone (PMIC). Speech Commun. 52 (April (4)).
Popović, B., Ostrogonac, S., Pakoci, E., Jakovljević, N., Delić, V., 2015. Deep neural network based continuous speech recognition for Serbian using the Kaldi toolkit. Lecture Notes in Computer Science: Speech and Computer, vol. 9319. pp. 186–192.
Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 2011. The Kaldi speech recognition toolkit. In: IEEE Workshop on Automatic Speech Recognition and Understanding, Hawaii, USA.
Rao, K.R., Yip, P., 1990. Discrete Cosine Transform: Algorithms, Advantages, Applications. Academic Press, Boston.
Rath, S.P., Povey, D., Vesely, K., Cernocky, J., 2013. Improved feature processing for Deep Neural Networks. In: Proceedings of Interspeech, Lyon, France.
Sahni, H., Bedri, A., Reyes, G., Thukral, P., Guo, Z., Starner, T., Ghovanloo, M., 2014. The tongue and ear interface: a wearable system for silent speech recognition. In: Proceedings of the 2014 ACM International Symposium on Wearable Computers, Seattle.
Salama, M., ElSherif, L., Lashin, H., Gamal, T., 2014. Recognition of unspoken words using electrode electroencephalograhic signals. In: Sixth International Conference on Advanced Cognitive Technologies and Applications, Cognitive 2014 (IARIA), Venice, Italy.
Schultz, T., Wand, M., 2010. Modeling Coarticulation in EMG-based continuous speech recognition. Speech Commun. 52 (April (4)), 341–353.
Stone, M., 2005. A guide to analysing tongue motion from ultrasound images. Clin. Linguist. Phonetics 19, 455–501.
Stone, M., Sonies, B., Shawker, T., Weiss, G., Nadel, L., 1983. Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system. J. Phonetics 11, 207–218.
Tang, L., Hamarneh, G., June 2010. Graph-based tracking of the tongue contour in ultrasound sequences with adaptive temporal regularization. Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE.
Tatulli, E., Hueber, T., 2017. Feature extraction using multimodal convolution neural networks for visual speech recognition. In: Proceedings of ICASSP2017, New Orleans, USA.
The Kaldi Toolkit, Online: http://kaldi-asr.org/.
Toth, A.R., Wand, M., Schultz, T., 2009. Synthesizing speech from electromyography using voice transformation techniques. In: Proceedings of Interspeech, Brighton, UK.
Tran, V.A., Bailly, G., Loevenbruck, H., Toda, T., 2010. Improvement to a NAM-captured whisper-to-speech system. Speech Commun. 52 (April (4)).
Wand, M., Himmelsbach, A., Heistermann, T., Janke, M., Schultz, T., 2013. Artifact removal algorithm for an EMG-based Silent Speech Interface. In: Proceedings of the IEEE Engineering Medicine and Biology Society Conference, pp. 5750–5753.
Wand, M., Schultz, T., 2011. Session-independent EMG-based speech recognition. In: International Conference on Bio-inspired Systems and Signal Processing, Proceedings of Biosignals, Rome, Italy.
Wand, M., Schultz, T., 2014. Pattern learning with deep neural networks in EMG-based speech recognition. In: Proceedings of EMBC14, IEEE, Chicago, USA, pp. 4200–4203.
Wang, J., Hahm, S., Mau, T., 2015. Determining an optimal set of flesh points on tongue, lips, and jaw for continuous silent speech recognition. In: 6th Workshop on Speech and Language Processing for Assistive Technologies, Dresden, Germany, 11 September.
Wang, J., Samal, A., Green, J.R., 2014. Preliminary test of a real time, interactive silent speech interface based on electromagnetic articulograph. In: Proceedings of the 5th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), Baltimore, Maryland USA, pp. 38–45.
Wang, J., Samal, A., Green, J.R., Rudzicz, F., 2012. Sentence recognition from articulatory movements for silent speech interfaces. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, 25-30 March, pp. 4985–4988.
Xu, K., Roussel, P., Gábor Csapó, T., Denby, B., 2017. Convolutional neural network-based automatic classification of midsagittal tongue gestural targets using B-mode ultrasound images. J. Acoust. Soc. Am. 141 (June (6)), EL531.
Xu, K., Yang, Y., Jaumard-Hakoun, A., Leboullenger, C., Dreyfus, G., Roussel-Ragot, P., Stone, M., Denby, B., 2015. Development of a 3D tongue motion visualization platform based on ultrasound image sequences. In: Proceedings of ICPhS2015, Glasgow.
Xu, K., Yang, Y., Stone, M., Jaumard-Hakoun, A., Leboullenger, C., Dreyfus, G., Roussel, P., Denby, B., 2016. Robust contour tracking in ultrasound tongue image sequences. Clin. Linguist. Phonetics 30 (3-5), 313–327.
Yamaguchi, H., Yamazaki, T., Yamamoto, K., Ueno, S., Yamaguchi, A., Ito, T., Hirose, S., Kamijo, K., Takayanagi, H., Yamanoi, T., Fukuzumi, S., 2015. Decoding silent speech in Japanese from single Trial EEGS: preliminary results. J. Comput. Sci. Syst. Biol. 8, 5.
Yamazaki, T., 2016. Silent Speech Brain-Computer Interface in Japanese. LAP Lambert Academic Publishing.
Yuksel, K.A., Buyukbas, S., Adali, S.H., 2011. Designing mobile phones using silent speech input and auditory feedback. In: Proceedings of MobileHCI 2011, August 30–September 2, Stockholm, Sweden.