Generating Intelligible Audio Speech from Visual Speech

Thomas Le Cornu, Ben Milner

***Abstract*—This work is concerned with generating intelligible audio speech from a video of a person talking. Regression and classification methods are proposed first to estimate static spectral envelope features from active appearance model (AAM) visual features. Two further methods are then developed to incorporate temporal information into the prediction - a feature-level method using multiple frames and a model-level method based on recurrent neural networks. Speech excitation information is not available from the visual signal, so methods to artificially generate aperiodicity and fundamental frequency are developed. These are combined within the STRAIGHT vocoder to produce a speech signal. The various systems are optimised through objective tests before applying subjective intelligibility tests that determine a word accuracy of 85% from a set of human listeners on the GRID audio-visual speech database. This compares favourably with a previous regression-based system that serves as a baseline which achieved a word accuracy of 33%.**

***Index Terms*—Audio-visual, speech reconstruction, DNN, RNN, STRAIGHT**

I. Introduction

THE aim of this work is to reconstruct an intelligible audio speech signal given only visual speech information taken from a speaker's mouth. Applications for such a system fall into the domain of silent speech interfaces which transform information from non-acoustic sensors into an audio signal and include electromagnetic articulography (EMA), ultrasound and electromyography (EMG) [1], [2], [3]. Silent speech interfaces have uses in a range of areas such as providing an artificial voice for patients who have undergone a laryngectomy [4]. Other areas include surveillance, where a video signal of a speaker is available but no audio is present, due possibly to the distance to the target or that no audio recording device is present. Conversely, providing privacy for cellular telephone users with silent speech input is a further application. To support real-time conversations, silent speech interfaces should operate sufficiently fast such that the "mouth-to-ear" delay is no more than 150ms [5].

The approach taken in this work is to exploit the correlation that exists between visual and audio speech [6], [7]. Several areas of speech processing have utilised this correlation by introducing a visual modality to provide complementary and robust features that are largely unaffected by audio noise. Some visual Lombard effects have been reported and can affect correlation [8]. Speech recognition systems have benefited from including visual features in low signal-to-noise ratios where reductions in word error rate have been reported [9], [10], [11]. Automatic machine lipreading systems rely completely on the information that can be extracted from the visual stream but have lower recognition accuracies in comparison to audio speech recognition [12], [13], [14]. In speech enhancement, visual speech information has provided spectral representations from which Wiener filters can be derived [15]. Additionally, speaker separation applications have reported benefits when extracting visual information from the speakers within the mixture [16]. Previous work on converting visual speech into audio has been limited, with our own earlier regression-based method as one example [17]. Transforming audio features to visual features for animation has also been studied, e.g. [18], which aims to produce realistic facial movement.

The proposed approach to reconstructing audio speech from visual speech is based on obtaining the parameters necessary to drive a model of speech production. Such a model requires acoustic features comprising vocal tract (spectral envelope) and excitation parameters, such as fundamental frequency, a measure of aperiodicity and a non-speech/unvoiced/voiced decision. In conventional speech coding applications these features are extracted from the audio signal, however, in this work only a visual speech representation is available. Intuitively, and from correlation analysis, it is apparent that certain acoustic features will not be available from the visual stream, such as excitation, while some level of spectral envelope information is present. Our previous work used Gaussian mixture models and deep neural networks within a regression framework to estimate spectral envelope features from visual features [17]. In this work we propose a second method for estimating audio features from visual features that uses a classification framework, whereby visual vectors are used to predict codebook entries that correspond to audio vectors. This approach is inspired by several speech processing systems that use codebooks to constrain estimates to be within a subset of all possible values. For example, inventory-style enhancement maps input noisy mel-frequency cepstral coefficient (MFCC) vectors to clean MFCC vectors via a joint codebook trained on clean and noisy speech [19]. Codebooks have also been used to provide maximum-likelihood estimates of speech and noise linear predictor parameters by searching for the combination of codebook entries that maximise the likelihood [20]. Furthermore, in this work we now place much greater importance on the dynamic nature of speech by considering long-range temporal information. Including temporal information for the regression and classification approaches is explored at the feature-level, by grouping contiguous frames of windowed speech, and at the model-level, using recurrent neural networks to model the relationship between sequences of input visual

and output audio features. Due to the limitations on audio speech information available from visual features, our goal is speech intelligibility rather than speech quality.

An alternative approach to create audio from video is to employ a visual-only speech recogniser (VSR) to extract a model sequence (e.g words) which is input into an audio text-to-speech (TTS) system [21]. Assuming the TTS is fully intelligible then the resulting word accuracy of the system would be dependent on the accuracy of the visual speech recogniser. Many silent speech interfaces adopt a similar approach and extract features from input sensors (e.g. EMA and EMG) which are then used to train a non-acoustic speech recogniser. In operation, data collected from the sensors is decoded by the speech recogniser into a word sequence which is input into a TTS system [2], [3]. Compared to the direct mapping from visual to audio proposed in this work, such a VSR/TTS approach would require a transcribed set of audio-visual data to train the visual speech recogniser and the audio TTS system. Conversely, in the proposed approach no such transcription is required as training is unsupervised. Additionally, and importantly for real-time operation, the VSR/TTS approach would likely have longer delay than the proposed direct approach. Even assuming that no language modelling is employed in the VSR, at least the end of each word spoken would need to be reached in the VSR decoding before the word could be synthesised. This would result in mouth-to-ear delays in excess of 150 ms which is considered to be an upper bound for conversational speech [5]. In addition, the audio may also suffer from jitter that is dependent on the variation of word lengths. As the proposed system outputs audio directly, the delay will be lower and dependent upon the duration and latency of input visual features

The remainder of the paper is organised as follows. Section II gives an overview of the speech reconstruction model, and the audio and visual speech features. Section III reviews the regression approach and describes the proposed classification method of spectral envelope estimation. Section IV then describes the feature-level and model-level approaches for including temporal information. Methods to generate aperiodicity and fundamental frequency are discussed in Section V. Objective and subjective experiments to measure intelligibility are presented in Section VI.

II. Speech reconstruction model and features

Speech production models typically require excitation and vocal tract (spectral envelope) features as input. The primary difficulty when attempting to reconstruct audio speech from visual speech is that the visual articulators provide information from only part of the vocal tract. In comparison, the audio speech signal is realised from the state of all the vocal organs [22]. Considering the features required by a speech production model, along with information that can be derived from visual featues, the following observations are made:

- fundamental frequency cannot be obtained as the vibration of the vocal folds cannot be seen;
- similarly, it is likely to be difficult to estimate a reliable voicing decision, especially when compounded with visual confusions of voiced and unvoiced phonemes [22];
- thus, the only parameter that could be estimated, albeit partially, relates to the vocal tract, or spectral envelope.

A. Speech model

Many speech production models have been proposed for coding and synthesis and include vocoders and harmonic models [23], [24]. Given its success in hidden Markov model (HMM) speech synthesis [21], and its ability to produce highly intelligible speech, this work uses the STRAIGHT vocoder [25]. This requires three acoustic features:

- a time-frequency spectral-envelope surface, $X(f, i)$;
- a fundamental frequency contour, f_0;
- a time-frequency measure of aperiodicity, $A(f, i)$;

where f and i represent the frequency bin and frame index respectively. As discussed previously, it is assumed that reliable excitation information cannot be estimated from the visual speech. Therefore, artificial values of fundamental frequency and aperiodicity are introduced to provide these missing parameters and are discussed in Section V. The audio features to represent vocal tract information and the visual features used to represent visual articulatory information are now discussed.

B. Audio features

Numerous features have been proposed to represent audio speech and include linear predictive coding (LPC) coefficients, MFCCs, perceptual linear prediction (PLP) coefficients and filterbank energies [26], [27]. Our previous work on visual to audio mapping [17] established that mel-filterbank energies are an effective audio feature to represent spectral envelope both for input to STRAIGHT and in mapping from visual features. Specifically, audio feature vectors, $\mathbf{a}_i$, are represented as mel-spaced log filterbank amplitudes which are extracted from the spectral envelope time-frequency surface produced by STRAIGHT analysis at rate of 100Hz. Preliminary investigations examined the effect of the number of filterbank channels and established that a dimensionality of $D^A = 22$ gave highest intelligibility. A sampling frequency of 8 kHz was used as our aim is to produce intelligible speech rather than high quality speech and observations found that audio-visual correlation deteriorates as frequency increases [28].

To obtain a spectral envelope surface from mel-filterbank features (such as may be estimated from visual features) that can be input into STRAIGHT, the following method is used

$$X(f, i) = \text{interp}\,(\mathbf{a}_i, \mathbf{f}_{\text{MEL}}) \quad (1)$$

where the interp function produces data points in the range 0-4 kHz given an input audio vector, $\mathbf{a}_i$, and set of mel-spaced frequencies, $\mathbf{f}_{\text{MEL}}$. Compensation is applied to equalise high-frequency spectral tilt that is introduced from the non-uniform bandwidths of the mel-spaced filterbank channels [29].

C. Visual features

Numerous features also exist for representing visual speech and include image-based and model-based methods. Image-based features, such as the two-dimensional discrete cosine transform [30] (2D-DCT) and sieve decomposition [31],

transform pixel intensities within an area located around the mouth to represent the appearance. Model-based features, such as the active shape model [32], use pre-trained statistical models to fit shape contours to the lips. Active appearance model [33] (AAM) based features have been particularly successful, outperforming other visual features in audio-visual speech recognition and lip reading tasks [34], and combine shape and appearance information. Consequently, in this work we extract $D^V = 13$ dimensional AAM features from the mouth region, as located by a mouth tracker from the video signal [35], to give visual feature vectors, $\mathbf{v}_i$. These are upsampled, using cubic interpolation, to match the frame rate of the audio vectors. Finally, both audio and visual features are z-score normalised.

III. Static spectral envelope estimation

This section describes two methods for estimating static filterbank vectors from visual speech. The first uses a regression approach [17]. The second uses a clustering and classification framework which first vector quantises a training set of audio vectors to create a codebook. A classification model then uses an input visual vector, $\mathbf{v}_i$, to predict a cluster label from the audio codebook which forms the audio filterbank estimate, $\hat{\mathbf{a}}_i$.

A. Regression method

Our previous work formulated estimation of audio vectors from visual vectors as one of regression [17], described by

$$\hat{\mathbf{a}}_i = g(\mathbf{v}_i) \tag{2}$$

where g is a fully-connected DNN architecture that is used to perform the mapping from input visual vector, $\mathbf{v}_i$, to output audio vector, $\mathbf{a}_i$. The network is trained using the backpropagation of errors algorithm in conjunction with stochastic gradient descent, and training is concluded once the validation error no longer decreases [36]. To find an optimal (or close to) set of hyperparameters a random search was performed over the number of hidden layers, the number of units within the layers, the learning rate and the dropout probability applied to hidden layers [37]. The model found to perform best consisted of three hidden layers each with 1024 units and uses the rectified linear unit (ReLU) activation function. For regularisation, dropout is applied after each hidden layer with probability $p = 0.5$. The final output layer uses linear activations and the network is trained with a learning rate of 0.0001.

B. Classification method

The second method reformulates estimation as one of classification where a model estimates a class label, $\hat{c}_i$, from the input visual feature, as

$$\hat{c}_i = o(\mathbf{v}_i) \tag{3}$$

where o is the classification model. Given the class label and a vector quantisation (VQ) codebook, C, the estimated audio vector, $\hat{\mathbf{a}}_i$ is output as

$$\hat{\mathbf{a}}_i = L(\hat{c}_i|C) \tag{4}$$

where the function L outputs (performing a simple lookup) the audio vector given the class label and codebook. This leads to a two-stage training requirement, to first create the VQ codebook and secondly the classification model.

1) VQ codebook training: The VQ codebook training uses only audio vectors. From a training set of N audio vectors, $\mathbf{X}^A = \{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$, the aim of the vector quantisation stage is to produce a codebook, C, of cluster centres with size K where $|C| = K$ and $K \ll \mathbf{X}^A$. The set of cluster centres, $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$, is found using the mini-batch k-means algorithm instead of the classic LBG algorithm as this is reported to converge more quickly to a comparable solution, given a large number of training samples [38].

Given the trained codebook, C, each audio feature vector, $\mathbf{a}_i$, in the training data, $\mathbf{X}^A$, is assigned a class label, c_i, by finding which cluster centre $\mathbf{c} \in C$ minimises the Euclidean distance and is obtained using

$$c_i = \underset{j}{\operatorname{argmin}} \|\mathbf{c}_j - \mathbf{a}_i\|^2 . \tag{5}$$

Having quantised each audio vector in $\mathbf{X}^A$, a corresponding set of labels is created for each vector, $\{c_1, \ldots, c_N\}$.

2) Classification model training: The classification model, o, in (3) is implemented using a DNN and is similar to that used for regression, described in Section III-A, except that the model now maps from input visual vector, $\mathbf{v}_i$, to the estimated codebook class label, $\hat{c}_i$, created during VQ training in (5). To obtain the output codebook class from the model, the softmax function is applied to the output of the final layer to give a set of class probabilities where the highest is selected. The same architecture, set of hyperparameters and training method as used for the regression model was found to give good performance for classification.

IV. Inclusion of temporal information

The previous section considered static audio vector estimation with no consideration of temporal information. It is well known that context is important in speech processing due to phenomena such as co-articulation and the inherent temporal structure [39]. This section extends estimation by proposing feature-level and model-level methods of incorporating temporal information into the estimation framework.

A. Feature-level temporal information

Instead of using single audio and visual vectors at each time instance, these are now grouped into audio and visual matrices, $\mathbf{A}_i$ and $\mathbf{V}_i$, that comprise S^A audio vectors and S^V visual vectors, respectively. The matrices contain an odd number of vectors, centred on the middle vector and defined as

$$\mathbf{A}_i = [\mathbf{a}_{i-w^A}; \ldots; \mathbf{a}_i; \ldots; \mathbf{a}_{i+w^A}] \tag{6}$$

$$\mathbf{V}_i = [\mathbf{v}_{i-w^V}; \ldots; \mathbf{v}_i; \ldots; \mathbf{v}_{i+w^V}] \tag{7}$$

where $w^A = \frac{S^A-1}{2}$ and $w^V = \frac{S^V-1}{2}$ and the semi-colon operator indicates concatenation of vectors. Larger matrix widths include greater levels of temporal information. When applied

to the regression method of Section III-A, the DNN maps input visual matrices, $\mathbf{V}_i$, to audio matrices, $\mathbf{A}_i$. When combined with the classification approach of Section III-B, the mini-batch k-means algorithm is applied to audio feature matrices to produce a codebook where each entry now represents a sequence of S^A static vectors. The DNN architectures and training are similar to that used for static input except now the input layer contains $D^V \times S^V$ units.

With the static-only method of visual to audio estimation each visual vector produced a corresponding audio vector. Now, with the output being a matrix of feature vectors, $\hat{\mathbf{A}}_i$, several different methods of converting these to a sequence of audio feature vectors exist. Specifically, three approaches are considered and illustrated in Figure 1:

1) Shift-by-1: the visual window is shifted by one vector at each time instance and the middle vector in the estimated audio feature matrix, $\hat{\mathbf{A}}_i$, is selected as the output.
2) Overlap-and-add: the visual window is shifted by one vector at each time instance and an audio matrix is output, $\hat{\mathbf{A}}_i$. These are then time-aligned, scaled by a triangular window and overlapped-and-added to form the output vector sequence.
3) Shift-by- S^A: the visual window is shifted by the size of the audio window, S^A, such that output audio matrices, $\hat{\mathbf{A}}_i$, are non-overlapping and the vectors within each are concatenated to form the output sequence.

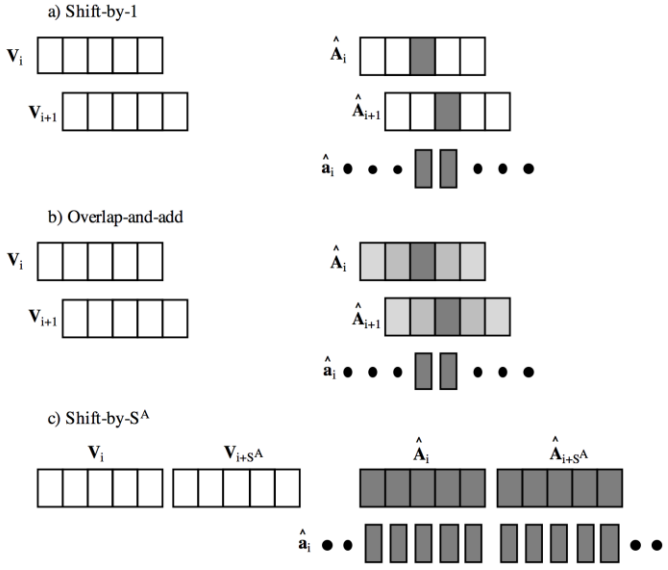

Fig. 1. Sliding window methods for converting estimated audio matrices to a sequence of audio vectors.

B. Model-level temporal information

Recurrent neural networks (RNNs) using the long short-term memory (LSTM) architecture have been successful in predicting output sequences from input sequences in a range of applications [40], [41]. These are now investigated as a further method to include temporal information into audio vector estimation. RNNs are an extension of standard neural networks and model dynamic processes by, in effect, introducing a feedback loop into the standard architecture. A sequence of T input visual vectors, $\{\mathbf{v}_1, \ldots, \mathbf{v}_T\}$, is passed through hidden layer weight connections to produce a hidden vector sequence, $\{\mathbf{h}_1, \ldots, \mathbf{h}_T\}$,

$$\mathbf{h}_t = \sigma\left(\mathbf{W}_{vh}\mathbf{v}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1}\right) \quad (8)$$

where σ is the ReLU activation function, $\mathbf{W}_{vh}$ are the input layer to hidden layer weights, and $\mathbf{W}_{hh}$ are the hidden to hidden layer weights. Bias terms have been omitted for clarity. Each element in the output sequence, $\{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ can be obtained through application of

$$\mathbf{y}_t = \mathbf{W}_{hy}\mathbf{h}_t \quad (9)$$

where $\mathbf{W}_{hy}$ are hidden layer to output layer weights. When configured for regression, each element in the output sequence is the output audio vector. For combination with classification, output vectors have K-dimensions that correspond to code-book class labels. To estimate an audio vector the softmax function is applied to each output vector and the codebook entry corresponding to the highest class probability is selected.

In the LSTM architecture the typical neural network units are replaced with memory cells. These cells store a value and use gates to control whether a new value is input, or if the current value should be output or forgotten. The benefit of using these specially designed units that store information is that they can exploit longer range dependencies present in the data. This behaviour is beneficial for speech processing ap- plications as it allows for modelling of dynamically changing context present in a time-varying signal such as speech [42].

Furthermore, typical RNN architectures only use past information to decide upon the current network output. However, it has been found that by also including future information the performance can be further improved over uni-directional models [43]. Bi-directional recurrent layers can be formed by using two hidden layers where one computes the forward hidden sequence $\overrightarrow{\mathbf{h}}$, and the other computes the reverse hidden sequence $\overleftarrow{\mathbf{h}}$. Element, $\mathbf{y}_t$, of the output sequence can then be obtained through application of

$$\mathbf{y}_t = \mathbf{W}_{\overrightarrow{h}y}\overrightarrow{\mathbf{h}}_t + \mathbf{W}_{\overleftarrow{h}y}\overleftarrow{\mathbf{h}}_t \quad (10)$$

To exploit the ability of deep neural network architectures to extract higher-level representations of the input data, multiple bi-directional LSTM layers can be stacked together. For our implementation this means the addition of another hidden layer, with processing performed backwards in time, such that the processing would begin at element T in the sequence, and work backwards to element one.

Following a search of hyperparameters using 4-fold cross-validation of the training data, a model with three hidden layers, each consisting of a forward and backward layer comprised of 250 LSTM units, was used. The final layer is either linear for the regression configuration or has the soft-max function applied for the classification approach. During training, regularisation is performed by introducing Gaussian noise to the inputs with a standard deviation of $\sigma = 0.6$ and by clipping gradients with values greater than one.

V. Aperiodicity and fundamental frequency

This section presents methods for producing excitation information. The STRAIGHT speech model requires a time-frequency aperiodicity surface and a fundamental frequency contour as inputs, in addition to spectral envelope. As discussed previously, we are unable to estimate aperiodicity and f_0 from the visual information. Therefore, methods have been developed to create artificial excitation information whilst maintaining as high a level of intelligibility as possible.

A. Aperiodicity

The aperiodicity surface, $A(f, i)$, represents the energy of aperiodic components in the frequency domain [44]. TTS systems that use STRAIGHT typically compress the aperiodicity surface into frequency bands [45]. Based on preliminary tests, five frequency bands are used in this work: 0-0.5 kHz, 0.5-1 kHz, 1-2 kHz, 2-3 kHz, and 3-4 kHz. From an input audio signal an aperiodicity vector, $\mathbf{p}_i$, is extracted

$$\mathbf{p}_i = \left[p_i^{(0-0.5)}, p_i^{(0.5-1)}, p_i^{(1-2)}, p_i^{(2-3)}, p_i^{(3-4)} \right]^{\cdot} \tag{11}$$

where $p_i^{(f_1-f_2)}$ is the aperiodic energy within the frequency band f_1 to f_2. Given a training set of aperiodicity vectors, these are combined with the respective audio filterbank vectors, $\mathbf{a}_i$, and the mini-batch k-means algorithm applied to create a joint filterbank-aperiodicity codebook, C^{A-P}. For aperiodicity estimation, given an estimated audio vector, $\hat{\mathbf{a}}_i$, the audio vector component of the codebook entries, $\mathbf{c}_j^A$, are searched and the aperiodicity component, $\mathbf{c}_{j^*}^P$, of the closest matching entry, j^*, is output as $\hat{\mathbf{p}}_i$, i.e.

$$\hat{\mathbf{p}}_i = \mathbf{c}^{P^*} \quad \text{where} \quad j^* = \underset{j}{\operatorname{argmin}} \left\| \mathbf{c}^{A} - \hat{\mathbf{a}}_i \right\|^2. \tag{12}$$

For input into STRAIGHT, cubic interpolation is applied to the five channels of $\hat{\mathbf{p}}_i$ to produce the aperiodicity surface, $A(f, i)$, using the procedure used in (1) for converting filterbank to spectral envelope. Aperiodicity estimation was analysed using codebook sizes from 1 to 512. Measuring the MSE of the resulting aperiodicity estimates found that a codebook comprising just 8 clusters gave lowest error. Codebooks with larger numbers of clusters were found to produce erratic aperiodicity surfaces that had higher MSE which adversely affected the intelligibility of the reconstructed speech.

B. Fundamental frequency

In our previous work we evaluated three methods for producing the artificial fundamental frequency contour parameter required by STRAIGHT [17]:

1) Unvoiced excitation: where all frames were assumed unvoiced.
2) Monotonic excitation: where the f_0 contour was set to a constant frequency.
3) Time-varying excitation: where the f_0 contour oscillated slowly about a mean frequency in an attempt to mimic the natural pitch fluctuations of spoken speech.

It was found that the unvoiced and monotonic methods performed best, as the time-varying method adversely affected the intelligibility of the resulting speech [46]. Considering this, and the result of further testing, we now use a monotonic f_0 contour as input to STRAIGHT. For male speech this value is set to 100Hz and for female speech to 207Hz, these being mean f_0 values obtained from training data.

VI. Experiments

The intelligibility of speech reconstructed from visual features is now evaluated. Experiments use the GRID audio-visual speech corpus that contains recordings from thirty-four speakers [47]. The corpus was chosen as the grammar has no context which makes subjective intelligibility testing unbiased. Each GRID sentence contains 6 words and follows the grammar in Table I. Each speaker produced 1000 utterances totalling around 50 minutes of speech per speaker. Speaker four (female) and speaker three (male) were used to create two speaker-dependent systems as they were found to be some of the most clear speakers within GRID [47]. Testing uses 200 sentences taken randomly from each speaker with training using the remaining 800 sentences.

TABLE I
GRID sentence grammar.

Command	Colour	Preposition	Letter	Digit	Adverb
bin	blue	at	A-Z	1-9	again
lay	green	by	minus W	zero	now
place	red	in			please
set	white	with			soon

The first tests use mean squared error (MSE) analysis to examine the effectiveness of estimating audio filterbank vectors and consider the inclusion of temporal information. Objective tests are then presented to predict the intelligibility and quality of speech reconstructed from estimated audio filterbank vectors in combination with the aperiodicty and fundamental frequency methods of Section V. Subjective evaluations are then performed using human listening tests.

A. Analysis of filterbank estimation using mean squared error

MSE analysis is used to first compare the three sliding window methods for producing an audio vector sequence and then to investigate inclusion of temporal information through feature-level and model-level methods.

1) Sliding window methods: Using the feature-level method of including temporal information, MSE tests are performed to compare the three sliding window methods proposed in Section IV-A and to examine the effect of the size, K, of the VQ codebook. The tests use audio information only and compute the MSE between the original un-quantised audio filterbank vectors and the resulting VQ output sequence of audio vectors. Figure 2 shows MSE for the three methods as K is increased from 16 to 4096 using an audio window width of $S^A = 31$. The results show consistently that overlap-and-add gives lower MSE. Other window width combinations were also investigated (but not shown) and had an identical

trend. Overlap-and-add does involve more computation than Shift-by-S^A but the smoothing it introduces is beneficial and generates less erratic trajectories between neighbouring frames. Therefore, the overlap-and-add method is adopted for all subsequent tests.

Increasing the codebook size reduces MSE quickly up to $K = 512$ which then decreases more slowly thereafter. Informal listening tests confirmed this trend and indicated that given a sufficiently large codebook size ($K \geq 512$) the resulting quantisation error was sufficiently small so that the intelligibility of reconstructed audio was indistinguishable from the original audio.

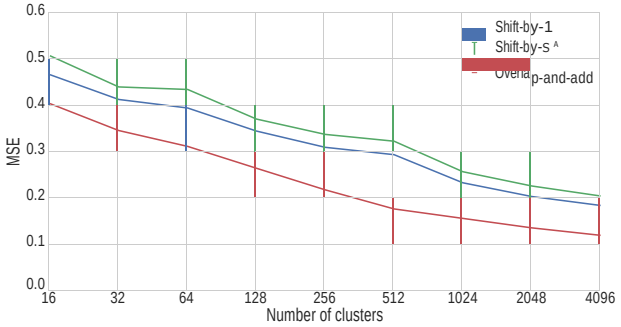

Fig. 2. Comparison of mean squared errors for the three sliding window techniques and for varying codebook sizes.

2) Feature-level temporal inclusion: Using the feature-level method of including temporal information, MSE analysis is performed to investigate how the sequence of audio filterbank vectors, estimated from visual feature matrices, changes as the amount of temporal information is varied through S^A and S^V. For each combination, a DNN is trained using codebook sizes of $K = \{512, 1024, 2048, 4096\}$ with results reported for the model that minimises error. Figures 3a and 3b show MSE for audio and visual window sizes of $S^A, S^V = \{3, 7, \ldots, 31, 35\}$ using the regression and classification approaches. For the classification method, MSE decreases rapidly as S^A and S^V are first increased and then stabilises at longer duration widths. Lowest MSE of 0.318 is attained with $S^A, S^V = 31$. For regression, MSE is more affected by increasing S_V while the audio window width has less effect. Lowest MSE of 0.325 is attained with $S^V = 35$ and $S^A = 23$, although several values of S^A produce similar MSEs.

3) Model-level temporal inclusion: MSE analysis is now applied to the model-level approach of including temporal information to investigate the effect of the sequence length, T, of input static visual vectors. Sequence lengths of $T = \{1, 3, 7, \ldots, 31, 35\}$ are examined for the regression and classification approaches and results shown in Figure 4. MSE reduces as sequence lengths increase and stabilises beyond $T = 11$. Both methods attain lowest MSEs at sequence lengths of 35 vectors, beyond which no reduction was observed, with regression having a lower MSE of 0.339 compared to 0.381 with classification. These are higher than reported using the feature-level method of including temporal information, although both approaches favour long duration widths in the region 310 ms to 350 ms. Incorporating such long duration

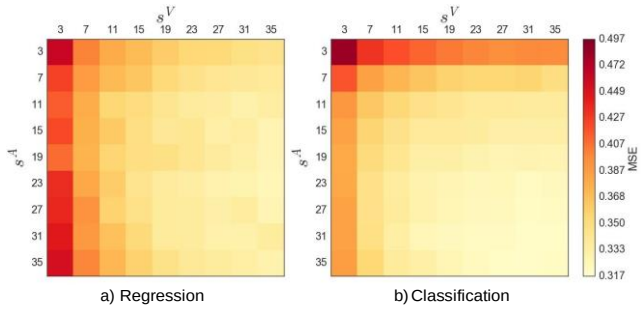

Fig. 3. Mean squared error of audio vectors estimated using a) regression and b) classification, with feature-level temporal information for varying audio and visual window sizes, using overlap-and-add.

windows of speech is supported by psychoacoustic studies of the peripheral human auditory system where it has been suggested that time spans of several hundred milliseconds of speech are integrated, as opposed to the short duration frames most commonly used in speech processing [48]. Additionally, ASR systems have shown benefits from incorporating temporal windows up to 1000 ms in length [49].

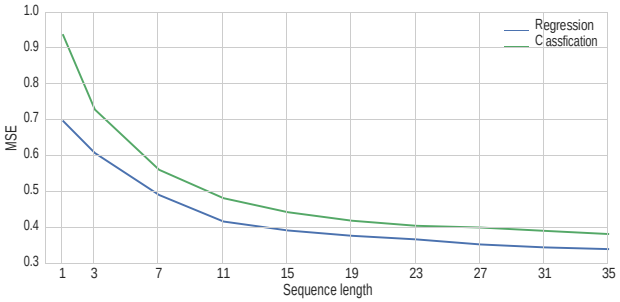

Fig. 4. Mean squared error of audio vectors estimated using regression and classification, with model-level temporal information for varying sequence lengths.

B. Objective testing

Objective measurements are now made of speech reconstructed using the regression and classification static estimates combined with the feature-level and model-level temporal approaches using ESTOI and PESQ [50], [51]. ESTOI is used to estimate the intelligibility of the reconstructed speech, while PESQ indicates the quality of the speech. Even though we state earlier that quality is not of primary concern, PESQ measurements provide a useful performance metric.

1) Objective intelligibility results – ESTOI: Tables II and III show ESTOI for female and male speakers, respectively, using the regression and classification approaches combined with feature-level temporal information for audio and visual window sizes of $S^A, S^V = \{23, 27, 31, 35\}$. ESTOI for regression and classification combined with model-level temporal information is shown in Table IV for female and male speakers using sequence lengths of $T = \{23, 27, 31, 35\}$.

Comparing regression and classification methods across feature-level and model-level methods of including temporal information, and across male and female speakers, shows

ESTOI to be consistently higher using the classification approach. With feature-level temporal information, ESTOI is slightly higher with wider visual window sizes and smaller audio window size, and for the classification method peaks at $S^A = 23$ and $S^V = 35$ for both female and male speakers. For regression, ESTOI peaks at $S^A = 27$ and $S^V = 35$. These equate to a broader visual context being used to estimate a narrower audio context. Using feature-level temporal information gives higher ESTOI compared to using model-level temporal information across male and female speakers and across regression and classification methods. For the classification method, with model-level temporal information, larger window widths improve ESTOI while for regression the performance is more variable. With no temporal information included, the regression and classification static features attain ESTOIs of 0.263 and 0.169 respectively for the male speaker and 0.201 and 0.200 for the female speaker, which are substantially lower than when temporal information is included.

TABLE II
ESTOI FOR FEMALE SPEAKER USING REGRESSION AND CLASSIFICATION APPROACHES WITH FEATURE-LEVEL TEMPORAL ENCODING.

	Regression				Classification			
S_A \ S_V	23	27	31	35	23	27	31	35
23	—	—	0.380	0.382	—	—	**0.434**	**0.434**
27	—	—	0.367	**0.384**	—	—	0.418	0.420
31	0.367	0.373	0.378	0.372	0.416	0.417	0.432	0.432
35	0.367	0.369	0.377	0.378	0.425	0.428	0.432	0.428

TABLE III
ESTOI FOR MALE SPEAKER USING REGRESSION AND CLASSIFICATION APPROACHES WITH FEATURE-LEVEL TEMPORAL ENCODING.

	Regression				Classification			
S_A \ S_V	23	27	31	35	23	27	31	35
23	—	—	0.397	0.401	—	—	0.435	**0.437**
27	—	—	0.402	**0.403**	—	—	0.423	0.424
31	0.373	0.387	0.387	0.396	0.418	0.424	0.422	0.422
35	0.388	0.384	0.392	0.390	0.431	0.431	0.431	0.431

TABLE IV
ESTOI USING REGRESSION AND CLASSIFICATION APPROACHES WITH MODEL-LEVEL TEMPORAL ENCODING FOR FEMALE AND MALE SPEAKERS.

	Female		Male	
T	Regression	Classification	Regression	Classification
23	0.339	0.373	0.371	0.407
27	0.337	0.379	0.377	0.414
31	0.334	0.379	**0.393**	0.415
35	**0.350**	**0.386**	0.381	**0.419**

2) Objective quality results – PESQ: PESQ analysis follows the same structure as the ESTOI analysis and uses the same set of configuration parameters. Tables V and VI show PESQ for the female and male speakers using feature-level temporal information combined with regression and classification methods. PESQ for model-based estimation is shown in Table VII. The PESQ results follow a similar trend to the ESTOI analysis. With feature-level temporal information, highest PESQ scores for both regression and classification are with visual windows, S^V, of 35 or 31 frames and associated audio window widths of between 4 and 12 frames smaller. PESQ for model-level temporal information again peaks at higher numbers of frames across male and females speakers and is lower than using feature-level temporal information.

The MSE analysis in Section VI-A and the ESTOI and PESQ analysis have shown the importance of including temporal information in the visual to audio estimation process. Highest performance comes with temporal widths of around 310ms to 350ms although differences in performance tend to decrease as window widths become wider. Bold text in the results tables shows best performance and those found using ESTOI form the settings for the subjective tests.

TABLE V
PESQ FOR FEMALE SPEAKER USING REGRESSION AND CLASSIFICATION APPROACHES WITH FEATURE-LEVEL TEMPORAL ENCODING.

	Regression				Classification			
S_A \ S_V	23	27	31	35	23	27	31	35
23	—	—	1.606	**1.620**	—	—	1.678	1.678
27	—	—	1.599	1.595	—	—	1.670	1.677
31	1.582	1.592	1.595	1.607	1.659	1.663	1.679	**1.686**
35	1.576	1.583	1.590	1.605	1.670	1.680	1.678	1.682

TABLE VI
PESQ FOR MALE SPEAKER USING REGRESSION AND CLASSIFICATION APPROACHES WITH FEATURE-LEVEL TEMPORAL ENCODING.

	Regression				Classification			
S_A \ S_V	23	27	31	35	23	27	31	35
23	—	—	1.949	1.958	—	—	2.052	**2.055**
27	—	—	**1.959**	1.952	—	—	2.031	2.025
31	1.902	1.921	1.930	1.932	2.017	2.022	2.023	2.025
35	1.915	1.909	1.931	1.924	2.027	2.030	2.038	2.038

TABLE VII
PESQ USING REGRESSION AND CLASSIFICATION APPROACHES WITH MODEL-LEVEL TEMPORAL ENCODING FOR FEMALE AND MALE SPEAKERS

	Female		Male	
T	Regression	Classification	Regression	Classification
23	1.555	1.549	1.900	1.759
27	1.573	1.552	1.909	1.823
31	1.574	1.554	**1.963**	1.822
35	**1.582**	**1.573**	1.939	**1.830**

C. Subjective testing

The objective tests have identified suitable configurations for the visual to audio reconstruction systems that are now analysed subjectively through human listening tests. The regression and classification static estimation approaches are combined with the feature-level and model-level methods of including

temporal information to give four systems namely: REG F, CLA F, REG M and CLA M. Furthermore, to investigate the usefulness of the original video these four methods are also tested when subjects are able to watch the video as well as hearing the audio. These represent audio-visual tests. A further test measures intelligibility when subjects are presented with just the video which is a lip-reading task, VID. Finally, as a baseline to our earlier work the regression only approach, REG, is also included [17]. This gives ten different systems which are summarised in the first four columns of Table VIII. Thirty-two subjects took part in the listening tests and each was presented with three sentences from each of the ten configurations, played in a random order. Subjects recorded the words they heard/saw through a web interface.

1) Listening test results: Table VIII shows the intelligibility (word accuracy) scores averaged across all subjects for the ten test configurations. These show that the proposed regression and classification approaches with temporal information (REG $\{$F,M$\}$ and CLA $\{$F,M$\}$) produce substantially higher intelligibility speech than the baseline regression method (REG). As found with the MSE analysis and objective tests, the subjective tests show the classification approach to produce more intelligible speech than the regression approach. Similarly, including temporal information using the feature-level method produces higher intelligibility speech than the model-level method. Highest overall intelligibility is achieved with the classification approach combined with feature-level temporal information (CLA F) which attains a word accuracy of 84.5%. Including the video signal in the tests improved intelligibility for the regression-based methods but not for classification approaches. However, intelligibility of the classification approaches is higher than for regression methods and it may be that the addition of video is unable to provide any further information. Intelligibility of the video (i.e. lip-reading (VID)) was substantially lower at 52.3%, although better than regression (REG). That said, all methods attained intelligibilities higher than chance, which is 19% for GRID.

Considering the alternative method of generating an audio signal by employing VSR followed by TTS (as discussed in Section I), visual-only word accuracies on GRID are between 80% and 97% [13], [14]. Assuming the subsequent TTS is fully intelligible then the lower end of this range is comparable to the proposed CLA F approach although the higher end is greater. It should be noted that these VSR systems are speaker-independent and moving to speaker-dependent VSR, to be equivalent to the CLA F method, would likely increase their accuracy. However, for real-time conversational speech, the delay using VSR/TTS will be at least the duration of each word, which Figure 5 shows to be longer than the 150ms recommended for conversational speech [5]. Conversely, in the proposed method, the latency is approximately half the visual window width, S^V, which is close to 150ms. In fact, latency can be reduced further by using non-symmetric windows that have fewer visual vectors ahead of the current time frame and more behind. Informal listening tests suggest that audio produced in this manner is different for an $S^V = 35$ frame window which had 8 visual vectors ahead of the current time point and 26 behind to give a latency of 90ms.

TABLE VIII
WORD ACCURACIES (AND STANDARD DEVIATIONS) FROM SUBJECTIVE LISTENING TESTS SHOWING THE INTELLIGIBILITY OF EACH CONFIGURATION.

Static	Temporal	Video	Name	Accuracy, %
Regression	None	None	REG	33.0 (11.7)
Regression	Feature	None	REG_F	77.3 (10.5)
Regression	Model	None	REG_M	72.2 (11.0)
Classification	Feature	None	CLA_F	84.5 (7.4)
Classification	Model	None	CLA_M	79.2 (7.5)
Regression	Feature	Video	REG_F_V	79.0 (11.8)
Regression	Model	Video	REG_M_V	77.1 (11.7)
Classification	Feature	Video	CLA_F_V	83.0 (6.5)
Classification	Model	Video	CLA_M_V	79.2 (6.8)
NA	NA	Video	VID	52.3 (16.2)

TABLE IX
PER-WORD ACCURACY SCORES FOR EACH OF THE SIX DIFFERENT SYSTEM CONFIGURATIONS.

	REG	REG F	REG M	CLA F	CLA M	VID
Command	28.1	89.6	91.7	97.9	94.8	57.3
Colour	58.3	94.8	95.8	100.0	100.0	75.0
Preposition	37.5	59.4	62.5	83.3	76.0	43.8
Letter	10.4	40.6	22.9	31.3	18.8	17.7
Digit	25.0	85.4	75.0	94.8	89.6	41.7
Adverb	38.5	93.8	85.4	100.0	95.8	78.1

2) Analysis of confusions: The GRID grammar in Table I shows that there are greater numbers of choices for letters (25) and digits (10) compared to the other categories which have four choices each (command, colour, preposition and adverb). To investigate this further, Table IX shows the word accuracy for each grammar category for the five audio-only configurations tested and for video-only. This reveals significant variation in word accuracy between the categories. Considering CLA F, as this performed best, word accuracy for command, colour and adverb categories approaches 100%, while for prepositions it is lower at 83% and for letters the accuracy is much lower at 31%. Clearly the number of choices within each category has an effect on word accuracy but we also speculate that word duration is a significant factor.

This is explored in Figure 5 which shows a scatter plot of the mean duration of each word in the GRID grammar (averaged over all occurrences of that word) against the mean word accuracy for that word in the subjective listening tests. Considering first the word accuracy of the preposition category, which is lower than that of the command, colour and adverb categories, these words have durations less than 200ms which is considerably shorter than the command, colour and adverb words which are, in general, longer. Letters, which have lowest accuracy, are all short with duration around 200ms. In fact, words with a duration over 300ms are all recognised with greater than 60% accuracy.

3) Spectrogram analysis: To illustrate the audio information extracted from the visual speech features, Figure 6 shows wideband spectrograms of the sentence '*Lay white with F 3 now*' taken from the female speaker, for the original

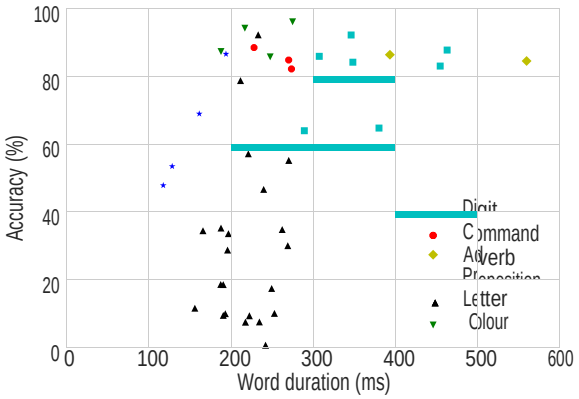

Fig. 5. Scatter plot of average word duration and word accuracy, broken down into GRID grammar categories.

signal and for those reconstructed using static regression (REG), and classification with feature-level (CLA F) and model-level (CLA M) temporal information. Confirming the objective and subjective test results, the feature-level and model-level spectrograms appear similar and show a significantly better representation of the formant structure of the original speech than static regression. One clear artefact of the reconstructed speech is widening of formant bandwidths which is a phenomena that also occurs in statistical parametric speech synthesis due to the averaging of frames with slightly different spectral structure [21]. Such averaging occurs when creating the VQ codebook and further examination of reconstructed speech shows the bandwidth of the first formant to exhibit relatively low broadening which then increases for higher frequency formants. Examples of speech reconstructed from visual features using the methods in this work are available at https://www.uea.ac.uk/computing/speech-language-and-audio-processing/v2a-results.

VII. Conclusions

This work has shown that intelligible audio speech can be generated solely from visual speech information. Specifically, a classification approach combined with feature-level temporal information using a deep neural network attained intelligibility of 85%. This outperformed a baseline regression method which had an intelligibility of 33%. Including temporal information was found to be important in producing intelligible speech with best performance attained with visual window widths of around 300ms in duration. This introduces a latency of around 150ms which is within the limit suggested to support real-time conversational speech. The alternative approach of combining visual speech recognition with TTS could attain intelligibility up to 97%, although the delay introduced by the decoding would be substantially longer, and dependent on word lengths, which is less practical for conversational speech.

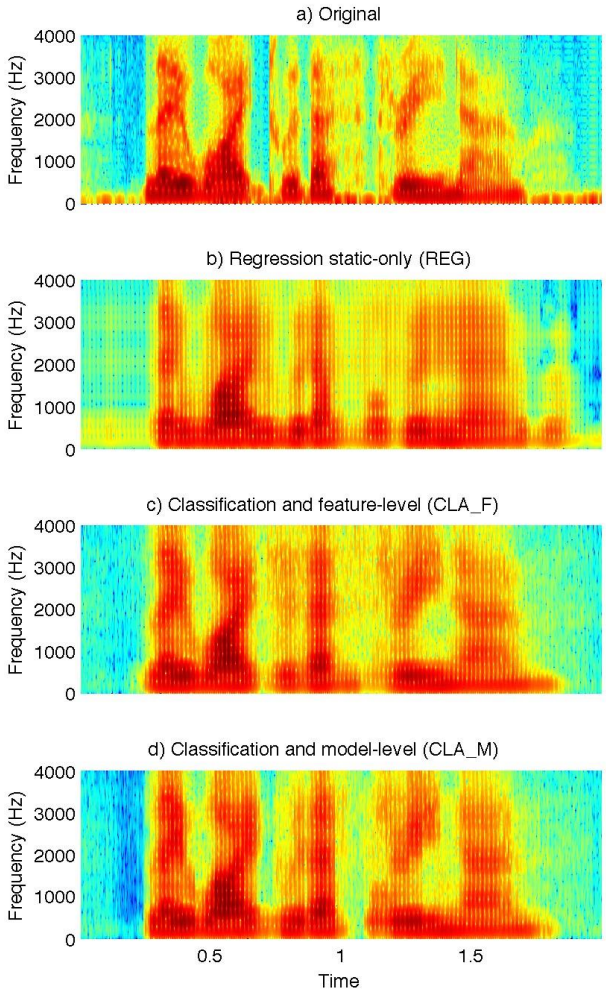

Fig. 6. Spectrograms of sentence '*Lay white with F 3 now*' spoken by the female speaker, showing, a) original speech, b) regression, c) classification with feature-level (CLA_F) and d) classification with model-level (CLA _M).

References

[1]B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, and J. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, Apr. 2010.

[2]C. Jorgensen and S. Dusan, "Speech interfaces based upon surface electromyography," *Speech Communication*, vol. 52, pp. 354–366, 2010.

[3]M. Fagan, S. Ell, J. Gilbert, E. Sarrazin, and P. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," *Medical Engineering and Physics*, vol. 30, no. 4, pp. 419–425, May 2008.

[4]J. A. Gonzalez, P. D. Green, R. K. Moore, L. A. Cheah, and J. M. Gilbert, "A non-parametric articulatory-to-acoustic conversion system for silent speech using shared gaussian process dynamical models," in *Fifth Speech Conference of UK and Ireland*, 2015.

[5]ITU-T, *G.114: SERIES G: One-way transmission time*. ITU-T recommendation, 2003.

[6]J. P. Barker and F. Berthommier, "Evidence of correlation between acoustic and visual features of speech," *ICPhS*, pp. 199–202, 1999.

[7]I. Almajai, B. Milner, and J. Darch, "Analysis of correlation between audio and visual speech features for clean audio feature prediction in noise." in *Interspeech*, 2006.

[8]S. Alexanderson and J. Beskow, "Animated lombard speech: Motion capture, facial animation and visual intelligibility of speech produced in adverse conditions," *Computer Speech & Language*, vol. 28, no. 2, pp. 607–618, 2014.

[9]G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *Issues in visual and audio-visual speech processing*, vol. 22, p. 23, 2004.

[10]H. Glotin, D. Vergyr, C. Neti, G. Potamianos, and J. Luettin, "Weighting schemes for audio-visual fusion in speech recognition," in *ICASSP*, vol. 1, 2001, pp. 173–176.

[11]S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous

speech recognition," *IEEE Trans. on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.
[12]Y. Lan, R. Harvey, and B.-J. Theobald, "Insights into machine lip reading," in *ICASSP*, 2012, pp. 4825–4828.
[13]Michael Wand and Jan Koutn ík and Jürgen Schmidhuber, "Lipreading with long short-term memory," in *ICASSP*, 2016.
[14]J. S. Chung, A. W. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," *CoRR*, vol. abs/1611.05358, 2016. [Online]. Available: http://arxiv.org/abs/1611.05358
[15]I. Almajai and B. Milner, "Visually derived Wiener filters for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1642–1651, 2011.
[16]F. Khan and B. Milner, "Speaker separation using visual speech features and single-channel audio." in *Interspeech*, 2013, pp. 3264–3268.
[17]T. Le Cornu and B. Milner, "Reconstructing intelligible audio speech from visual speech features," in *Interspeech*, 2015, pp. 3355–3359.
[18]C. Luo, J. Yu, X. Li, and Z. Wang, "Realtime speech-driven facial animation using Gaussian mixture models," in *Multimedia and Expo Workshops (ICMEW)*, July 2014, pp. 1–6.
[19]X. Xiao and R. Nickel, "Speech enhancement with inventory style speech resynthesis," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1243–1257, Aug. 2010.
[20]S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 163–176, Jan. 2006.
[21]H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.
[22]L. Bernstein, "Visual speech perception," *AudioVisual Speech Processing*, pp. 21–39, 2012.
[23]R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.
[24]Y. Stylianou, J. Laroche, and E. Moulines, "High-quality speech modification based on a harmonic + noise model," in *Eurospeech*, 1995, pp. 451–454.
[25]H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," in *ICASSP*, 2008, pp. 3933–3936.
[26]H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.
[27]ETSI, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm," ETSI STQ-Aurora DSR Working Group, ES 202 212 version 1.1.1, Nov. 2003.
[28]I. Almajai and B. Milner, "Maximising audio-visual speech correlation," in *AVSP*, 2007.
[29]B. Milner and X. Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 24–33, 2007.
[30]I. Almajai and B. Milner, "Using audio-visual features for robust voice activity detection in clean and noisy speech," in *EUSIPCO*, 2008, pp. 1–5.
[31]I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213, 2002.
[32]T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models - their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.
[33]T. F. Cootes, G. J. Edwards, C. J. Taylor *et al.*, "Active appearance models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.
[34]Y. Lan, R. Harvey, B. Theobald, E.-J. Ong, and R. Bowden, "Comparing visual features for lipreading," in *AVSP*, 2009, pp. 102–106.
[35]E. Ong, Y. Lan, B. Theobald, R. Harvey, and R. Bowden, "Robust facial feature tracking using selected multi-resolution linear predictors," in *ICCV*, 2015.
[36]G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
[37]J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 281–305, 2012.
[38]D. Sculley, "Web-scale k-means clustering," in *Proc. WWW*. ACM, 2010, pp. 1177–1178.
[39]W. Holmes, *Speech synthesis and recognition*. CRC press, 2001.
[40]A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2009, pp. 545–552.
[41]A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013, pp. 6645–6649.
[42]H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." in *Interspeech*, 2014, pp. 338–342.
[43]A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *ASRU*, 2013, pp. 273–278.
[44]H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight." in *MAVEBA*, 2001, pp. 59–64.
[45]H. Silen, E. Helander, and M. Gabbouj, "Prediction of voice aperiodicity based on spectral representations in HMM speech synthesis." in *Interspeech*, 2011, pp. 105–108.
[46]D. Websdale, T. Le Cornu, and B. Milner, "Objective measures for predicting the intelligibility of spectrally smoothed speech with artificial excitation," in *Interspeech 2015*, 2015.
[47]M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.
[48]S. Sharma, D. Ellis, S. Kajarekar, P. Jain, and H. Hermansky, "Feature extraction using non-linear transformation for robust speech recognition on the aurora database," in *ICASSP*, vol. 2, 2000, pp. 1117–1120.
[49]B. Y. Chen, Q. Zhu, and N. Morgan, "Learning long-term temporal features in LVCSR using neural networks." in *Interspeech*, 2004.
[50]J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE Trans. Audio, Speech and Language Processing*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
[51]R. Rix, J. Beerands, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment telephone networks and codecs," in *ICASSP*, 2001, pp. 749–752.

Thomas Le Cornu was born in Birmingham, UK. He received a BSc degree in Computing Sciences in 2013, and a PhD degree in audiovisual speech processing in 2017, both from the University of East Anglia, UK. His current work combines computer vision, machine learning, and electronic engineering for producing intelligent sensor devices and analytics software for biotechnology applications. His research interests include audiovisual speech processing, pattern recognition, and high-throughput biotechnology systems.

Ben Milner was born in Norfolk, UK. He received a BEng degree in electronic engineering in 1991 and a PhD degree in signal processing in 1995, both from the University of East Anglia, UK. From 1994 he spent seven years working at BT Laboratories, specialising in noise and channel robustness and feature extraction for speech recognition. In 2001 he moved to the School of Computing Sciences at UEA where his research interests include speech enhancement, audio-visual processing and machine learning.