Reconstructing intelligible audio speech from visual speech features

Thomas Le Cornu and Ben Milner

University of East Anglia
{t.le-cornu, b.milner}@uea.ac.uk

Abstract

This work describes an investigation into the feasibility of producing intelligible audio speech from only visual speech features. The proposed method aims to estimate a spectral envelope from visual features which is then combined with an artificial excitation signal and used within a model of speech production to reconstruct an audio signal. Different combinations of audio and visual features are considered, along with both a statistical method of estimation and a deep neural network. The intelligibility of the reconstructed audio speech is measured by human listeners, and then compared to the intelligibility of the video signal only and when combined with the reconstructed audio.

Index Terms: speech intelligibility, visual speech, GMMs, DNNs, STRAIGHT

1. Introduction

The aim of this project is to produce intelligible audio speech solely from visual speech features extracted from the mouth region of a speaker. This is motivated by a desire to be able to listen to speech from a speaker when no audio is present. Our specific application is a surveillance scenario where only the video of a speaker is available. Other scenarios also exist where a speaker is too far away from a microphone or when no audio channel is available.

Using visual speech information within speech processing applications that have traditionally used only the audio signal has received significant interest in recent years. One of the earliest applications of combining audio and visual speech information is in audio-visual speech recognition [1, 2]. Including visual speech information can give a significant improvement in recognition accuracy in noisy conditions where the visual features are largely (visual Lombard effects have been reported [3]) insensitive to the noise that distorts the audio features. However, in noise-free conditions the gain is at best only marginal as the visual features tend not to provide complementary information to the audio features. Automatic lip reading relies completely on the visual features to decode the speech signal and consequently reports much lower recognition accuracy than systems that also use audio features. In one study, decoding accuracy was 42% with visual-only features in comparison to 87% with audio-visual features [4], which indicates the limit of visual speech information. Visual features have also been used effectively in audio speech enhancement where they are transformed into an audio spectral representation to form an enhancement filter (e.g. Wiener filter) that is applied to the input noisy speech [5, 6]. Results have shown the enhanced speech to be largely free from the original noise signal but to be rather distorted. The related area of speaker separation has also benefited from visual speech information taken from the speakers in the mixture [7, 8]. This has been used to construct binary masks for separation and to address permutation and scaling ambiguities present after blind source separation.

To synthesise an audio speech signal, a speech model requires a number of acoustic speech features that would normally be extracted during an analysis stage. A typical set of acoustic features comprises a voiced/unvoiced/non-speech classification, fundamental frequency (for voiced speech), spectral envelope, and phase. It is clear that a visual-only signal will lack many of these parameters. However, several studies have shown correlation to exist between audio and visual speech features, the magnitude of which depends on the specific features being tested [9, 10, 11]. For example, low-resolution spectral envelope audio features have higher correlation than more detailed spectral features. Conversely, excitation parameters, such as fundamental frequency have no correlation to visual features, although it may be possible to infer some indication of voicing from whether teeth are visible or not. Visual voice activity detection (VAD) has been more successful with several schemes proposed that outperform audio-only VADs at lower signal-to-noise ratios (SNRs) [12]. These factors make reconstructing audio speech from only visual speech features a challenging problem. Considering the source-filter components of speech, it can largely be assumed that no voicing or fundamental frequency information can be inferred, while only a broad spectral envelope can be estimated with limited accuracy.

Our approach to this problem is to estimate a smoothed spectral envelope from the visual features to provide vocal tract information. No attempt is made to estimate voicing or fundamental frequency and instead a number of methods for producing an artificial excitation signal are considered. Given this set of speech parameters, a model of speech production is then used to reconstruct a time-domain audio signal. As so many components of the speech signal are effectively missing, our metric for success is the intelligibility of the reconstructed speech signal rather than its quality.

The remainder of the paper is organised as follows. Section 2 provides a brief overview of the speech reconstruction model and then explains how voicing and fundamental frequency are determined artificially. Section 3 describes the audio and visual speech features that have been considered. A discussion and comparison of the estimators used is given in Section 4. Section 5 describes the results and analysis of experiments that first examine objectively the effectiveness of audio feature estimation from visual features and secondly, using human listening tests, determines the intelligibility of the reconstructed speech.

2. Speech reconstruction model

Many models of speech production have been proposed for speech coding and synthesis applications. These include vocoders, sinusoidal models and harmonic-plus-noise models [13, 14]. Based on successful application in hidden Markov model (HMM) synthesis to produce intelligible speech, the STRAIGHT vocoder has been chosen [15, 16].

STRAIGHT is a sophisticated implementation of a channel vocoder that separates speech into its spectral envelope and source components and was developed to allow for flexible manipulation of parameters to produce high-quality speech modifications. To synthesise a time-domain speech signal STRAIGHT requires three sets of parameters: the fundamental frequency, f_{0_i}; a measure of aperiodicity, $A(f,i)$; and a time-frequency surface, $X(f,i)$; where i and f represent the frame index and frequency bin.

When considering reconstructing a speech signal from only visual speech information it is not possible to estimate several of these parameters. Neither the fundamental frequency nor the aperiodicity can be estimated from the visual speech features. Instead suitable values need to be produced artificially. As the key performance metric is intelligibility and not quality, suitable values will need to be found that maximise intelligibility, even though this may lead to poor quality. The time-frequency surface is the only parameter than can be estimated from the visual speech features.

2.1. Aperiodicity and fundamental frequency

Three "artificial" methods of setting the voicing and fundamental frequency are considered. The first method, *monotone*, sets the fundamental frequency for each frame to a constant value, i.e $f_{0_i} = \mu_{f_0}$, which produces monotone sounding speech. From f_0 analysis of the speaker used in testing (discussed in Section 5) it was found that $\mu_{f_0} = 216\,\text{Hz}$. The second method, *time-varying*, modulates the monotone f_0 contour using a 0.25 Hz sinusoid with an amplitude that gives a frequency change, Δ_{f_0}, of ± 28 Hz.

$$f_{0_i} = \mu_{f_0} + \Delta_{f0} cos((2\pi i/400) + \phi_r) \tag{1}$$

where ϕ_r is a random phase offset. The settings for the time-varying parameters were established by examining fundamental frequency contours of real speech and synthesising speech to follow the trends measured. The final method, *unvoiced*, synthesises an entirely unvoiced speech signal and uses Gaussian white noise as the excitation. For unvoiced excitation, all time-frequency aperiodicity values, $A(f,i)$, are set to zero while in voiced speech they are set to $-\infty$.

A further method was also included in testing, *original*, and uses the original voicing and fundamental frequency estimated from the time-domain speech signal using PRAAT [17]. Whilst this is not realistic in real operating conditions it provides a useful baseline for evaluation.

2.2. Spectral envelope

Visual speech features exhibit correlation with spectral envelope which gives the possibility of estimating the time-frequency surface, $X(f,i)$, required by STRAIGHT, from the visual information, i.e.

$$\hat{X}(f,i) = g(\boldsymbol{v}_i) \tag{2}$$

where $\boldsymbol{v}_i$ is the visual vector and g is the estimator. The next two sections consider choices for the audio and visual features and then two methods of estimation.

3. Audio and visual speech features

Combinations of audio and visual features are considered with the aim of identifying combinations that enable audio features to be estimated from visual features with low error.

3.1. Audio features

Two spectral envelope representations are considered: linear prediction coding (LPC) coefficients and mel-filterbank

3.1.1. LPC

LPC analysis is a common technique for estimating vocal-tract filter coefficients. From each frame of speech, LPC analysis is applied to create an LPC vector, $\boldsymbol{a}_i$. Different filter orders, P, were considered, specifically with $P = \{2,4,6,8,14\}$, where lower orders introduce more smoothing into the spectral envelope. Synthesising speech with $P = 14$ produced speech that was almost indistinguishable from the original speech signal.

3.1.2. Filterbank

Filterbanks are frequently used in speech processing applications to reduce spectral detail into a spectral envelope-like representation. This work uses a mel-filterbank based on that specified in the ETSI Aurora standard [18]. The number of filterbank channels, K, was varied with $K = \{4,7,10,15,20\}$. The log of the resulting filterbank channel amplitudes was taken but no further processing was applied. Speech reconstructed from the 20-channel filterbank gave best sounding speech and was comparable to the LPC-14 configuration.

3.2. Visual features

Visual features can be model-based or pixel-based, and both have been applied successfully in audio-visual speech processing [6, 19, 20]. This analysis considers a model-based feature, the active appearance model (AAM); and a pixel-based feature, based on a 2D discrete cosine transform (2D-DCT).

3.2.1. 2D-DCT

Two-dimensional DCT features are extracted from a 128×128 matrix of pixel intensities, $\boldsymbol{P}$, that is centred on a tracked mouth centre point and resampled. A 2D-DCT is applied to produce coefficient matrix, $\boldsymbol{C}$, from which a visual vector, $\boldsymbol{v}_i^{2DDCT}$, is obtained by extracting coefficients in a zigzag order from the lower coefficient region of the matrix [21] to give a J-dimensional visual vector

$$\boldsymbol{v}_t^{2D-DCT} = [c_{0,0}, c_{0,1}, c_{1,0}, c_{2,0}, c_{1,1}, ...] \tag{3}$$

where $c_{m,n}$ are elements of $\boldsymbol{C}$. Preliminary tests found best estimation performance when 36 coefficients were retained and when combined with velocity temporal derivatives.

3.2.2. AAM

AAM features are commonly used in audio-visual speech processing to model shape and appearance [19]. Shape parameters are a concatenation of the coordinates of the set of vertices detailing the outline of the inner and outer mouth, while appearance parameters are pixel intensities extracted from the mesh of the current visual frame that has been warped to the base shape. From a set of training images, each annotated with landmark points that outline features of the lip contours and eyes, the AAM warps each image to a mean shape and appearance and then builds a statistical model using principal component analysis (PCA). PCA is applied to shape and appearance separately and then to the concatenated shape and appearance feature vectors to produce a compact visual feature.

Given a test image, the AAM minimises the difference between its synthesised face image and the actual face image by varying model parameters as well as incurring displacements in position, scale, and orientation. Features of the final synthesised image form a J-dimensional AAM vector, $\boldsymbol{v}_t^{AAM}$. Preliminary tests found best performance with 13 AAM coefficients, augmented with their velocity temporal derivatives.

4. Audio feature estimation

Given a sequence of visual vectors, $\boldsymbol{v}_i$, extracted from a speaker, the task of estimation is to identify a corresponding sequence of audio vectors, $\hat{\boldsymbol{a}}_i$, which can then be transformed into the time-frequency surface, $\hat{X}(f,i)$, needed by STRAIGHT to synthesise a time-domain speech signal. Two forms of estimator have been considered, one being a statistical estimator based on a GMM and the other a deep neural network (DNN).

4.1. Gaussian mixture models

Estimation of an audio feature vector, $\hat{\boldsymbol{a}}_i$, begins by creating a Gaussian mixture model (GMM) that models the joint density of the audio and visual feature vectors from a speaker. A joint feature vector, $\boldsymbol{z}_i$, is first created by augmenting audio vectors and visual vectors

$$\mathbf{z}_i = [\boldsymbol{a}_i, \boldsymbol{v}_i] \tag{4}$$

From a training set of joint feature vectors, expectation-maximisation (EM) clustering is applied to create a GMM, Φ^{av}, that models the joint density of the audio and visual features

$$\Phi^{av} = \sum_{c=1}^{C} \gamma_c \phi_c(\boldsymbol{z}) = \sum_{c=1}^{C} \gamma_c \, \mathcal{N}(\mathbf{z}; \mu^c, \boldsymbol{\Sigma}^c) \tag{5}$$

The GMM comprises C clusters with the cth cluster having a prior probability, γ_c; Gaussian probability density function, ϕ_c, with mean vector, $\boldsymbol{\mu}_c$, and covariance matrix, $\boldsymbol{\Sigma}_c$.

Given the model of the joint density of audio-visual vectors, Φ^{av}, an estimate of the audio vector, $\hat{\boldsymbol{a}}_i$, can be made from the visual vector extracted from the speaker's mouth region, $\boldsymbol{v}_i$,

$$\hat{\boldsymbol{a}}_i = \arg\max_{\boldsymbol{a}} \left(p\left(\boldsymbol{a}_i | \boldsymbol{v}_i, \Phi^{av}\right)\right). \tag{6}$$

4.2. Deep neural network

Recently, deep neural networks (DNN) have shown success in acoustic modelling for automatic speech recognition tasks, outperforming state-of-the-art GMM-HMM systems [22, 23]. In these configurations, the DNNs are being used for classification, that is, the prediction of HMM states given an input. In this paper, their use is explored for regression.

The DNN architecture used for this paper consists of a fully connected network with three hidden layers between the input layer and output layers. The three hidden layers each have 1024 units. The hidden units use Rectified Linear Units (ReLU) as the activation functions, and linear units were used in the output layer. The neural network was trained using resilient backpropagation [24] with mini-batches of 500 training examples. The learning rate was fixed at 0.001 and z-score normalisation was performed on the data. Network training was stopped once the R^2 of the test data stopped improving.

5. Experimental results

The aim of the experiments is to establish whether intelligible audio speech can be reconstructed from just visual speech features. Tests begin by first examining objectively how well audio features can be estimated from visual features using the GMM and DNN. Two of the best performing configurations then form the basis of subjective tests where human listeners are used to determine the intelligibility (word accuracy) of the reconstructed speech.

The GRID audio-visual database is used for the experiments [25]. Sentences comprise six words and follow the grammar shown in Table 1. Thirty-four speakers form the GRID database, with each speaker producing 1000 sentences. Speaker S4 (female) was chosen for the tests as the speech was considered to be articulated clearly and in a study of word accuracy across the speakers scored highly [25]. From the 1000 sentences, 800 are used for training and 200 for testing.

Table 1: 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

5.1. Objective measurements

An objective analysis is performed to determine how accurately the audio features are estimated from visual features using the GMM and DNN. The correlation between the estimated audio feature and that extracted directly from the original speech signal is measured. Tables 2 and 3 show correlation values for estimating LPC coefficients and filterbank amplitudes, using DNN and GMM classifiers, and AAM and 2D-DCT visual features.

Table 2: Correlation values, r, for LPC configurations.

	DNN		**GMM**	
Num. coeffs	**AAM**	**2D-DCT**	**AAM**	**2D-DCT**
2	0.59	0.62	0.73	0.72
4	0.61	0.62	0.72	0.71
6	0.57	0.59	0.72	0.72
8	0.62	0.65	0.73	0.71
14	0.52	0.53	0.71	0.72

Table 3: Correlation values, r, for filterbank configurations.

	DNN		**GMM**	
Num. channels	**AAM**	**2D-DCT**	**AAM**	**2D-DCT**
4	0.82	0.81	0.79	0.81
7	0.83	0.82	0.79	0.81
10	0.83	0.82	0.81	0.81
15	0.82	0.82	0.81	0.81
20	0.82	0.82	0.81	0.81

Comparing the audio features, filterbank amplitudes have substantially higher correlation to visual features than LPC coefficients across all configurations. Table 3 shows that the correlation of filterbank amplitudes is largely unaffected by the

Table 4: Methods of reconstructing audio speech from visual features

Method	TF surface	Excitation
GMM_ORIG	GMM + AAM + LPC	Original
GMM_UNV	GMM + AAM + LPC	Unvoiced
DNN_ORIG	DNN + 2D-DCT + Filterbank	Original
DNN_UNV	DNN + 2D-DCT + Filterbank	Unvoiced

choice of visual feature and estimator. Considering LPC coefficients, using a GMM for estimation outperforms the DNN in all cases. There is also little difference in correlation with respect to the visual feature used.

5.2. Intelligibility tests

The aim of the subjective intelligibility experiments is threefold. First, to examine whether reconstructing audio speech from visual features can produce intelligible speech. Second, to compare the intelligibility of the reconstructed audio with the intelligibility from just the video of the speaker, i.e. lip reading. Third, to examine whether combing reconstructed audio with the video improves intelligibility. To address these questions the subjects are presented with samples from three different multi-media configurations: the reconstructed audio-only, the original video only, and the reconstructed audio combined with the original video.

To generate the reconstructed audio four different configurations are examined. Two methods of estimating the time-frequency surface were used:

- GMM with AAM and 8th order LPC audio features
- DNN with 2D-DCT 20-channel filterbank audio features

These represent a small subset of the configurations analysed in Section 5.1, but it would be prohibitive to include all combinations in the listening tests. Instead, our approach is to use two very different configurations to examine their impact on intelligibility. These were combined with two methods for creating the speech excitation – using the original voicing and fundamental frequency and using fully unvoiced excitation. Again, it would be prohibitive to try all combinations of excitation in the listening tests, so preliminary tests determined that the unvoiced excitation gave the most intelligible audio of the three methods introduced in Section 2.1. These two choices of excitation allow the impact of having no knowledge of the voicing/fundamental frequency to be compared to having full knowledge. The four methods are summarised in Table 4.

XX listeners took part in the tests, which were carried out in a quiet environment with subjects using headphones and positioned in front of a monitor. Each subject was played (in a random order to remove any bias) 12 audio-only sentences, 12 audio-video sentences and 3 video-only sentences. The 12 audio sentences comprised 3 examples from each of the four configurations in Table 4. This gave a total of 27 sentences, all of which were different. Each listener was allowed to replay the audio/video as many time as they wished before entering the words they heard. This was done as a potential application of the work would be to transcribe speech from recordings where listeners would be able to replay recordings multiple times.

Table 5 shows the intelligibility (word accuracy) for the four different methods of reconstructing audio, listed in Table 4, when hearing just the reconstructed audio and when combined with the original video of the sentence. The intelligibility obtained using only the video was 49.02%. For the GRID grammar shown in Table 1, the intelligibility that would be expected by chance alone is 19% assuming unbiased test conditions.

Table 5: Intelligibility of reconstructed audio-only and audio-video speech.

Method	Audio-only	Audio-video
GMM_ORIG	48.37 %	60.46 %
GMM_UNV	40.20 %	53.27 %
DNN_ORIG	37.25 %	54.25 %
DNN_UNV	28.76 %	45.10 %

5.3. Discussion

The results show that both configurations which have no knowledge of the original audio (GMM_UNV DNN_UNV) are able to reconstruct audio speech from visual features with intelligibility higher than chance. When these are supplemented by the original video signal, the intelligibility increases further. Intelligibility with GMM_UNV audio and video is higher than using only the video and agrees with studies that show that an audio-visual signal is more intelligible than a single modality. The intelligibility of DNN_UNV audio and video remains lower than video only and is attributed to the lower intelligibility of the audio – around 12% lower than with GMM_UNV-based audio. Identifying the reason for this difference is not straightforward as the two configurations differ in their audio and visual features as well as the method of estimation. However, listening informally to speech produced by a range of different configurations suggests that the audio feature is most important when considering intelligibility, rather than the visual feature or method of estimation. The spectral envelope produced from estimated LPC coefficients is closer to the original spectral envelope than that produced by the filterbank dues to its relative coarseness.

Reconstructing audio using the fundamental frequency and voicing estimated from the original speech this gives an absolute increases in intelligibility of around 8% over using a purely unvoiced excitation. This demonstrate the important of voicing and is attributed to several of the vocabulary items requiring voicing to be classified correctly, such as /s/ and /z/ confusions.

6. Conclusions

This work has shown that it is possible to reconstruct an intelligible audio speech signal from just visual speech features. Compared to articulatory speech synthesis, where knowledge of articulators such as the tongue is available, the information from the video is limited to the mouth shape appearance, with no knowledge of the excitation signal available. For the purpose of estimating a time-frequency surface from the visual speech, LPC audio features have been found to be better as they are better able to model spectral surface.

Subjective tests found that the excitation provides important information for intelligibility. The reconstructed audio was found to have similar intelligibility to using just the video from the speaker (i.e. lip reading) but when subjects were presented with both the intelligibility increased. This task required of listeners is acknowledged to be rather constrained at present and important further work will extend the intelligibility tests to less constrained tasks and concentrate on optimising the audio features.

7. References

[1] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *Issues in Audio-Visual Speech Processing, MIT Press*, 2004.

[2] V. Estellers, M. Gurban, and J.-P. Thiran, "On dynamic stream weighting for audio-visual speech recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1145–1157, May 2012.

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

[4] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri, "Large-vocabulary audio-visual speech recognition: a summary of the Johns Hopkins Summer 2000 Workshop," in *IEEE Fourth Workshop on Multimedia Signal Processing*, 2001, pp. 619–624.

[5] L. Girin, J.-L. Schwartz, and G. Fang, "Audio-visual enhancement of speech in noise," *Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 3007–3020, Jun. 2001.

[6] A. Almajai and B. Milner, "Visually derived Wiener filters for speech enhancement," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 6, pp. 1642–1651, Aug. 2011.

[7] F. Khan and B. Milner, "Speaker separation using visual speech features and single-channel audio," in *Interspeech*, 2013.

[8] Q. Liu, W. Wang, and P. Jackson, "Audio-visual convolutive blind source separation," in *Sensor Signal Processing for Defence (SSPD 2010)*, 2010.

[9] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal tract and facial behavior," *Speech Communication*, vol. 26, no. 1, pp. 23–43, Oct. 1998.

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

[11] 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*, USA, Sep. 2006, pp. 2470–2473.

[12] I. Almajai and B. Milner, "Using audio and visual features for robust voice activity detection in clean and noisy speech," in *EUSIPCO*, Switzerland, Aug. 2008.

[13] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[14] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 1, pp. 21–29, Jan. 2001.

[15] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, Apr. 1999.

[16] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-independent HMM-based speech synthesis system – HTS-2007 system for the Blizzard Challenge 2007," in *Proc. Blizzard Challenge 2007*, Aug. 2007.

[17] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," 2001.

[18] ETSI, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms," ETSI STQ-Aurora DSR Working Group, ES 202 050 version 1.1.1, Oct. 2002.

[19] T.F.Cootes, G. Edwards, and C.J.Taylor, "Active appearance models," *IEEE Trans. PAMI*, vol. 23, no. 6, pp. 691–685, Jun. 2001.

[20] G. Meyer, J. Mulligan, and S. Wuerger, “Continuous audio-visual digit recognition using N-best decision fusion,” *Information Fusion*, vol. 5, no. 2, pp. 91–101, Jun. 2004.

[21] K. Sayood, *Introduction to Data Compression*. Morgan-Kaufmann, 2000.

[22] 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,” *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[23] G. E. Dahl, T. N. Sainath, and G. E. Hinton, “Improving deep neural networks for lvcsr using rectified linear units and dropout,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8609–8613.

[24] M. Riedmiller and H. Braun, “A direct adaptive method for faster backpropagation learning: The rprop algorithm,” in *Neural Networks, 1993., IEEE International Conference on*. IEEE, 1993, pp. 586–591.

[25] 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. 150, no. 5, pp. 2421–2424, Nov. 2006.