Using visual speech information and perceptually motivated loss functions for binary mask estimation

Danny Websdale, Ben Milner

University of East Anglia

d.websdale@uea.ac.uk, b.milner@uea.ac.uk

Abstract

This work is concerned with using deep neural networks for estimating binary masks within a speech enhancement framework. We first examine the effect of supplementing the audio features used in mask estimation with visual speech information. Visual speech is known to be robust to noise although not necessarily as discriminative as audio features, particularly at higher signal-to-noise ratios. Furthermore, most DNN approaches to mask estimate use the cross-entropy (CE) loss function which aims to maximise classification accuracy. However, we first propose a loss function that aims to maximise the hit minus false-alarm (HIT-FA) rate of the mask, which is known to correlate more closely to speech intelligibility than classification accuracy. We then extend this to a hybrid loss function that combines both the CE and HIT-FA loss functions to provide a balance between classification accuracy and HIT-FA rate of the resulting masks. Evaluations of the perceptually motivated loss functions are carried out using the GRID and larger RM-3000 datasets and show improvements to HIT-FA rate and ESTOI across all noises and SNRs tested. Tests also found that supplementing audio with visual information into a single bimodal audio-visual system gave best performance for all measures and conditions tested.

Index Terms: HIT-FA, speech separation, binary mask

1. Introduction

Speech separation from a monaural source aims to separate target speech from interfering background noise to produce a more intelligible signal. Such systems have widespread application in areas such as speech enhancement, robust speech recognition and hearing aid design [1, 2, 3]. There are two main approaches to this problem. The first is to derive a statistical model that makes certain assumptions about the background noise, and includes methods such as spectral subtraction, Weiner filtering and mean-square error estimation [4]. These approaches have been shown to not provide an increase in intelligibility for human listeners [5, 6]. This is because distortions (e.g. musical noise) are introduced and low-intensity sounds (e.g. unvoiced consonants), which are important for intelligibility, are lost. The second approach uses computational auditory scene analysis (CASA) [7], inspired by perceptual principles of auditory scene analysis (ASA), and can be effective in both stationary and non-stationary noise [8].

In CASA, speech is extracted by applying a mask to a time-frequency (T-F) representation of noisy speech. An ideal binary mask (IBM) retains speech dominant T-F units and suppresses noise dominant T-F units. An IBM can be constructed from premixed speech and noise and defined as

$$\text{IBM}(t,f) = \begin{cases} 1, & \text{if SNR}(t,f) \geq \text{LC} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

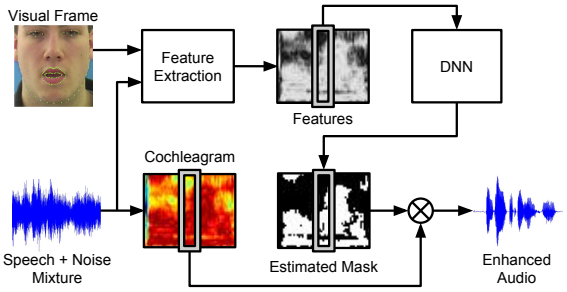

Figure 1: *Overview of the speech separation system.*

where t and f represent time frame and frequency bin respectively and LC is a local criterion. T-F units dominated by speech are assumed to have a signal-to-noise ratio (SNR) greater than or equal to LC are represented by 1 and retained. Noise dominant units are assumed to be less than LC are represented by 0 and suppressed. Several studies have reported subjective test results where IBMs improve intelligibility for speech in noise for both normal-hearing and hearing-impaired listeners [9, 10, 11, 12]. In practice an IBM is not available and instead the binary mask must be estimated from the noisy signal. This allows speech separation to be treated as a mask estimation problem that uses supervised learning to map acoustic features extracted from noisy speech to a binary mask [13].

This work considers two extensions to mask estimation. First, information from visual speech is investigated as a supplement to the audio information used in mask estimation. The use of visual speech information in traditionally audio-only speech processing applications has given significant gains in performance in noisy conditions. For example, in speech recognition, supplementing the audio with visual features has reduced error rates in low SNR conditions [14, 15, 16]. A benefit of using visual speech information within mask estimation is that visual features are not degraded by acoustic noise, although in themselves they may not have the discriminative ability that audio features possess in terms of mask estimation. To investigate this we explore mask estimation, and subsequently speech intelligibility, by comparing audio-only, visual-only and audio-visual speech mask estimation.

The second extension to mask estimation proposed in this work is development of perceptually motivated loss functions within a deep neural network (DNN) framework. Most existing methods of mask estimation using DNNs use the classification accuracy of the T-F units as the basis of the cross-entropy (CE) [17] loss function that is used during training [18, 19, 20]. However, several studies have shown that the HIT-FA rate of the mask correlates more closely to speech intelligibility than classification accuracy [18, 19, 20, 21, 22]. Therefore, we propose using perceptually motivated loss functions that are based on maximising the HIT-FA rate with the aim of increasing the intelligibility of the resulting masked speech.

Figure 1 shows the overall speech separation system. Features are extracted from noisy speech and visual frames and input into a DNN to estimate a binary mask. Masking is applied to a cochleagram [7] of the noisy speech which suppresses noise-dominated T-F units and the remaining signals are overlapped and summed to produce the enhanced signal. The same system is used for all speech enhancement configurations, except the visual stream is removed for audio-only, and the audio stream is removed for visual-only.

The remainder of the paper is organised as follows. Section 2 provides an overview of acoustic and visual feature extraction methods. The classifier and proposed loss functions are described in Section 3. Performance evaluations are made in Section 4 which first compare the performance of including visual information through visual-only and audio-visual systems, and secondly the effectiveness of the proposed perceptual loss functions under varying noise and SNR conditions using both a small dataset (GRID) and large dataset (RM-3000).

2. Audio-visual feature extraction

Feature extraction aims to identify suitably discriminative information in the noisy input speech and the visual speech that enables the DNN to determine whether T-F units are target (1) or noise (0) dominated.

2.1. Acoustic feature extraction

The acoustic feature selected is the multi-resolution cochleagram (MRCG) feature, designed specifically for cochleagram mask estimation. The MRCG feature combines four cochleagrams at different resolutions [20]. The first captures high resolution localised detail while the remaining cochleargrams capture lower resolution spectrotemporal content. Cochleagrams are computed by passing the input signal through a 64-channel gammatone filterbank [23].

The outputs from the gammatone filterbank are split into 20 ms frames with 10 ms frame shift with power spectrum computed followed by a log which gives the first cochleagram, CG_1. Similarly, CG_2 uses 200 ms frames with 10 ms frame shift. Finally, CG_3 and CG_4 are derived by applying an 11×11 and 23×23 mean filter kernel to CG_1 [20]. The final MRCG feature, $\mathbf{A}_t$, is produced by stacking all four CGs for time t.

2.2. Visual feature extraction

The visual feature selected is the active appearance model (AAM) which has proven to be an effective feature within audio-visual speech recognition [14, 24, 25] and is a model-based combination of shape and appearance. AAMs require labelled data with landmarks to generate features and use a model to perform this task automatically. The model requires hand labelled training images to learn the variation in mouth shapes and in this work 43 training images were used with 101 landmarks tracked. Forty-six and 20 landmarks represent the outer and inner lip respectively, with the extra landmarks for the eyes and jaw line, which assist the model in locating the face and fitting landmarks. A new model is produced by selecting only the mouth landmarks, and is used to produce AAM features, $\mathbf{v}_t = [\mathbf{s}_t\ \mathbf{a}_t]$, that comprise shape, $\mathbf{s}_t$, and appearance, $\mathbf{a}_t$, components for time t.

The shape feature, $\mathbf{s}$, is obtained by concatenating n x and y coordinates that form a two-dimensional mesh of the mouth, $\mathbf{s} = (x_1 y_1, ..., x_n y_n)^T$. A model that allows linear variation in shape is produced using PCA,

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{m} p_i \mathbf{s}_i \tag{2}$$

where $\mathbf{s}_0$ is the base shape, $\mathbf{s}_i$ are the shapes corresponding to the m largest eigenvectors and p_i are shape parameters. Coefficients comprising 90 % of the variation are selected, resulting in a vector size of 8 shape coefficients, $\mathbf{s}_t$.

The appearance feature, $\mathbf{a}$, is obtained from the pixels that lie inside the base mesh, $\mathbf{s}_0$ [26]. As with the shape model, an appearance model, $\mathbf{a}$, can also be expressed with linear variation,

$$\mathbf{a} = \mathbf{a}_0 + \sum_{i=1}^{m} q_i \mathbf{a}_i \tag{3}$$

where $\mathbf{a}_0$ is the base appearance, $\mathbf{a}_i$ are the appearances that correspond to the m largest eigenvectors and q_i are appearance parameters. Coefficients comprising 95 % of the variation are selected, giving a vector size of 15 appearance coefficients, $\mathbf{a}_t$. Combining the shape and appearance features gives an AAM vector, $\mathbf{v}_t$, with 23 dimensions which is extracted from the video at a rate of 25fps.

2.3. Temporal information

Including temporal information along with static features has shown to improve accuracy in automatic speech recognition (ASR) [27, 28]. In this work we include temporal information via vector stacking. Given a sequence of static feature vectors, $\{\dots, \boldsymbol{x}_{t-2}, \boldsymbol{x}_{t-1}, \boldsymbol{x}_t, \boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+2}, \dots\}$, neighbouring vectors within a window that extends K vectors either side of the current vector are stacked, i.e.

$$\boldsymbol{x}_t^{\mathrm{STACK}} = [\boldsymbol{x}_{t-K}, \dots, \boldsymbol{x}_t, \dots, \boldsymbol{x}_{t+K}] \tag{4}$$

Preliminary tests found a window of 7 frames (i.e. K=3) gave best performance. Due to the difference in frame rates between acoustic and visual features, visual features are upsampled to that of the acoustic features. For audio-only systems $\boldsymbol{x}_t = \mathbf{a}_t$, for visual-only $\boldsymbol{x}_t = \mathbf{v}_t$ and for audio-visual: $\boldsymbol{x}_t = [\mathbf{a}_t\ \mathbf{v}_t]$.

3. Perceptually motivated loss functions

The purpose of the classifier is to learn a mapping between the input audio and visual features extracted from the noisy speech and the binary mask output. Previous studies have shown a progression in classifiers used, beginning with GMMs through to support vector machines (SVMs), multilayer perceptrons (MLPs) and finally deep learning [19, 18, 20, 21, 22, 29]. We use DNNs as the classifier in this work which normally use the binary cross-entropy (CE) loss function in training for classification tasks. The DNN uses rectified linear units for hidden layers and a sigmoid layer for the output. The CE loss function is now reviewed and two new perceptually motivated loss functions introduced inspired by the HIT-FA rate.

3.1. Binary cross-entropy (CE) loss function

Binary cross-entropy (CE) is a standard loss function used within DNN training for classification tasks [17] and forms the baseline loss function. The aim of CE is to maximise the accuracy of the estimated mask where accuracy is defined as the proportion of correctly labeled T-F units. The CE loss, L^{CE}, is

calculated as

$$\mathrm{L}^{\mathrm{CE}} = -\frac{1}{N}\sum_{n=1}^{N}\Big[y_n \log(\hat{y}_n) + (1-y_n)\log(1-\hat{y}_n)\Big] \quad (5)$$

where $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ are vectors that comprise concatenated frames of T-F units for the true and predicted masks respectively, for each mini-batch in DNN training, from the IBM and estimated mask respectively. Each of these vectors comprises N T-F units which are indexed by n.

3.2. HIT-FA (HF) loss function

Our first perceptually motivated loss function (HF) is based on maximising the HIT-FA rate, which several studies have shown correlates more closely to intelligibility than mask accuracy [18, 19, 20, 21, 22]. In terms of the loss function, HITs refer to the proportion of correctly labeled target-dominant T-F units while FAs refer to the proportion of incorrectly labeled noise-dominant T-F units. Studies have shown that achieving high HITs and low FAs produces higher intelligibility [18, 19].

The key difference between the CE and HF loss functions is that CE calculates accuracy over all T-F units together, whereas HF calculates the accuracy of target-dominant (1) and noise-dominant (0) T-F units separately. HIT-FA has a range between 1 and -1, with 1 being best performance. However within DNN training loss is minimised, therefore we use FA-HIT to give best performance at -1 and remove this discrepancy. The HIT-FA loss, L^{HF}, is calculated as

$$\mathrm{L}^{\mathrm{HF}} = \frac{1}{S}\sum_{n=1}^{N}\Big[(1-y_n)\hat{y}_n\Big] - \frac{1}{R}\sum_{n=1}^{N}\Big[y_n\hat{y}_n\Big] \quad (6)$$

where S is the number of T-F units within $\boldsymbol{y}$ that should be suppressed (0s) and R is the number of T-F units within $\boldsymbol{y}$ that should be retained (1s).

3.3. Cross-entropy HIT-FA hybrid (CHF) loss function

Within an IBM the number of retained T-F units, R, and number of suppressed units, S, are generally not equal. In most cases there are more noise-dominant T-F units than target-dominant units, due mainly to areas of non-speech. The HF loss function is calculated as proportions of R and S separately, and is therefore less affected by bias towards a difference between R and S.

Conversely, the CE loss function is calculated as an overall accuracy of R and S and is therefore biased towards the greater of the two. We take inspiration from the HF loss function to produce a hybrid cross-entropy HIT-FA (CHF) loss function by modifying the CE function to remove this bias. To do this we normalise the ratio between R and S such that $R = S$. This is achieved by multiplying the portion related to S by R/S. The cross-entropy HIT-FA loss function, $\mathrm{L}^{\mathrm{CHF}}$, is calculated as

$$\mathrm{L}^{\mathrm{CHF}} = -\frac{1}{N}\sum_{n=1}^{N}\Big[y_n \log(\hat{y}_n) + \frac{R}{S}(1-y_n)\log(1-\hat{y}_n)\Big] \quad (7)$$

Our data has a bias towards S, therefore this normalisation will cause an increase of HITs at a cost of increasing FAs. The opposite would occur if the bias was towards R. A reduction to overall classification accuracy will occur in all cases where $R \neq S$ prior to normalisation.

4. Experimental results

The experiments examine the effect of including visual speech features in mask estimation through audio-only, visual-only and audio-visual tests. Comparisons are also made of the proposed loss functions across these different input feature configurations. Tests first use a small dataset (GRID) and then expand to a larger vocabulary dataset (RM-3000).

The GRID audio-visual dataset contains recordings from 34 speakers who each produced 1000 sentences [30]. Each sentence comprises six words and follows the grammar shown in Table 1. Speaker 12 (male) was selected for the speaker dependent evaluations with the audio downsampled to 16 kHz, and the video was captured at 25 fps. The speech database is split into 200 test sentences, and 800 training sentences of which 160 are removed for validation within training.

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

The second audio-visual dataset, RM-3000 [31], consists of 3000 sentences spoken by a single native English speaking male speaker. The sentences were randomly selected from the 8000 sentences in the Resource Management (RM) Corpus [32]. The vocabulary size of 1000 words and no strict grammar give a more realistic environment, and more challenging task when compared to GRID. The audio was downsampled to 16 kHz and the video was captured at 25 fps. The speech database is split into 600 test sentences, and 2400 training sentences of which 480 are removed for validation within training.

The DNN architecture for all experiments consisted of 5 hidden layers with 1024 rectified linear units [33] with an output sigmoid layer of 64 units. Input data was z-score normalised and grouped into minibatches of 256. Dropout of 0.2 was used after all hidden layers to prevent overfitting [34]. To further prevent overfitting, early stopping [35] was used within training for when the validation score had not improved after 10 further epochs, up to a maximum of 250 epochs. Training was performed using backpropagation with the RMSprop optimiser [36]. A learning rate of $3e^{-5}$ was chosen with 0.9 momentum applied. The DNN was built using the Lasagne [37] library with a Theano [38] backend.

For evaluating the performance of our speech separation systems, we utilise three objective measures: i) classification accuracy of T-F units between true (y) and predicted ($\hat{y}$) masks, ii) HIT-FA rate and iii) ESTOI [39]. Within the ESTOI function, non-speech frames are removed via dynamic range thresholding, however in our experiments, we found this method to perform poorly and not remove the desired non-speech frames. Therefore, we remove the non-speech frames using the alignment transcriptions provided for each dataset prior to the ESTOI function.

4.1. Analysis with GRID dataset

These tests use the GRID dataset to compare the performance of our proposed perceptual loss functions using audio-only, visual-only and audio-visual system. Experiments are performed in babble and factory noise at SNRs of -5 dB, 0 dB and +5 dB,

with LC set to be 5 dB lower than the selected SNR as this was found to give best performance in initial tests and conforms to that described in [21, 22]. Tables 2 and 3 show the classification accuracy, HIT-FA and ESTOI performance across babble and factory noise respectively for audio, visual and audio-visual systems using the CE, HF and CHF loss functions.

Table 2: *Classification accuracy (in %), HIT-FA (in %) and ESTOI scores for the GRID dataset in babble noise at* -5 dB, 0 dB *and* +5 dB.

SNR	Feat	Loss	Acc	HIT-FA (FA)	ESTOI
-5 dB	A	CE	**89.7**	66.7 (4.6)	**46.9**
		HF	84.8	68.0 (14.5)	42.6
		CHF	88.3	**71.7** (9.5)	46.1
	V	CE	**87.1**	63.0 (7.8)	**46.2**
		HF	84.6	**70.1** (15.8)	44.2
		CHF	85.5	68.7 (13.5)	45.9
	AV	CE	**91.0**	73.4 (5.2)	**53.7**
		HF	86.5	74.3 (14.2)	48.3
		CHF	89.4	**78.0** (10.2)	52.0
	unprocessed audio				20.3
0 dB	A	CE	**91.8**	74.7 (4.1)	62.4
		HF	88.7	77.4 (11.2)	60.3
		CHF	90.6	**79.5** (8.7)	**62.8**
	V	CE	**87.1**	62.7 (7.6)	53.1
		HF	84.5	**69.7** (15.7)	53.1
		CHF	85.4	68.3 (13.5)	**53.9**
	AV	CE	**92.1**	76.7 (4.5)	**64.8**
		HF	88.7	78.6 (11.9)	62.0
		CHF	90.8	**81.1** (9.0)	**64.8**
	unprocessed audio				33.9
+5 dB	A	CE	**92.6**	77.6 (4.0)	72.2
		HF	90.1	81.0 (10.3)	72.6
		CHF	91.4	**82.6** (8.5)	**74.1**
	V	CE	**87.1**	62.9 (7.8)	59.7
		HF	84.7	**69.6** (15.4)	62.5
		CHF	85.3	68.7 (13.9)	**62.9**
	AV	CE	**92.6**	78.2 (4.2)	72.6
		HF	89.4	80.9 (11.4)	72.4
		CHF	91.5	**82.8** (8.5)	**74.5**
	unprocessed audio				49.8

Table 3: *Classification accuracy (in %), HIT-FA (in %) and ESTOI scores for the GRID dataset in factory noise at* -5 dB, 0 dB *and* +5 dB.

SNR	Feat	Loss	Acc	HIT-FA (FA)	ESTOI
-5 dB	A	CE	**92.8**	69.1 (2.7)	**44.8**
		HF	89.4	74.1 (9.4)	40.9
		CHF	91.1	**75.7** (7.2)	43.8
	V	CE	**90.2**	64.4 (5.5)	44.1
		HF	87.3	**73.9** (12.5)	43.0
		CHF	88.6	71.9 (9.9)	**45.0**
	AV	CE	**93.5**	75.0 (3.3)	**50.7**
		HF	89.9	79.1 (10.0)	46.3
		CHF	91.9	**81.5** (7.5)	50.6
	unprocessed audio				20.1
0 dB	A	CE	**94.4**	76.9 (2.5)	58.7
		HF	91.3	79.9 (8.0)	57.2
		CHF	92.9	**83.2** (6.4)	**60.1**
	V	CE	**90.2**	64.7 (5.5)	50.4
		HF	87.5	**74.0** (12.2)	51.7
		CHF	88.6	72.2 (10.0)	**52.3**
	AV	CE	**94.5**	78.6 (2.7)	60.7
		HF	91.2	81.9 (8.7)	58.9
		CHF	92.9	**84.8** (6.9)	**62.3**
	unprocessed audio				33.5
+5 dB	A	CE	**95.1**	80.3 (2.4)	66.9
		HF	92.1	83.9 (7.9)	68.1
		CHF	93.6	**86.6** (6.2)	**70.6**
	V	CE	**90.3**	64.4 (5.4)	55.7
		HF	87.4	**74.4** (12.4)	**60.3**
		CHF	88.5	72.3 (10.2)	59.8
	AV	CE	**95.0**	81.1 (2.6)	67.8
		HF	91.6	83.9 (8.5)	68.6
		CHF	93.3	**87.0** (6.8)	**71.2**
	unprocessed audio				49.9

Firstly, comparing the performance of audio-only, visual-only and audio-visual systems across all noise types and SNRs, we find that all systems provide large gains in intelligibility over unprocessed audio. When combining audio and visual information into a bimodal system highest performance is found across all measures for all noise types and SNRs. Highest gains in performance over audio-only is found at low SNRs, where the visual information complements best the degraded audio. Gains of 8.3 and 5.8 in HIT-FA rate and gains of 6.8 and 5.9 for ESTOI over audio-only at -5 dB for babble and factory noise respectively are achieved, producing an overall improvement of 33.4 and 30.6 in intelligibility over the unprocessed audio. At high SNRs, the benefit gained from combining audio and visual information over audio-only is reduced as the audio features are less degraded by noise which allows the DNN to more effectively map to the target masks in these less challenging conditions.

For visual-only systems, classification accuracy and HIT-FA rate provides a consistent score across all SNRs for each noise type. This is due to the visual feature being unaffected by noise type or SNR corrupting the audio stream, and the performance is provided by how well the DNN can map the input visual features to the target mask. The only difference between noise type and SNR configurations are the configuration dependant target masks. The increase in intelligibility with increasing SNR through ESTOI is due solely to the less corrupted noisy mixtures at higher SNR.

Comparing now the effect of the loss functions with respect to classification accuracy, the CE loss function gives highest accuracy across all SNRs and noise types and across all configurations. This is expected as the CE loss function is targeted to maximise accuracy. The hybrid CHF loss function has accuracy almost as high as CE and exceeds that of HF which is not designed to maximise classification accuracy.

Considering now the HIT-FA rate, the HF loss function now outperforms the CE loss function as it is designed to maximise HIT-FAs. However, the hybrid CHF loss function gives even higher HIT-FAs across all SNRs and noise types for systems containing audio information, while the visual-only system consistently has highest HIT-FA rate with the HF loss function. In terms of HITs, the CHF and HF loss functions perform similarly, but their main difference is that the CHF loss function generates fewer FAs compared to the HF loss function. Lowest HITs and FAs are found with the CE loss function due to it favouring 0s over 1s in the mask, which is caused by the bias towards the larger of S and R. The CHF loss function is able to remove this bias and provides a balance between increasing HITs without increasing as many FAs.

Comparing now the intelligibility as measured by ESTOI, the CE loss function outperforms the HF loss function at lower SNRs while the HF loss function is better at the higher 5 dB SNR for all systems, and is better above 0 dB for visual-only systems. Even though the HF loss function outperforms CE with regards to the HIT-FA rate across all configurations, the large number of FAs introduced by the HF loss function reduces the intelligibility to be lower than CE at low SNRs. This shows that even a large increase in HITs does not compensate for a large increase in FAs, which are more detrimental to intelligibility at low SNR than at high SNR. Considering now the performance of the hybrid CHF loss function, this outperforms both CE and HF at SNRs above -5 dB and is slightly worse than CE at -5 dB. The CHF loss function had higher HIT-FA rate over CE across all SNR for all systems, confirming that increasing the HIT-FA rate does increase intelligibility, but the number of FAs introduced affects the resulting intelligibility. Reducing FAs at low SNRs is critical whereas a higher HIT rate is more important at high SNRs.

Overall, with intelligibility being the main focus, all systems provide large gains in ESTOI over unprocessed audio, with the bimodal audio-visual system outperforming both audio-only and visual-only across all configurations. With regards to loss functions, if the SNR is very low, CE is the loss function of choice, however at all other SNRs, CHF is the best performing loss function. CHF also provides a strong balance between both classification accuracy and the HIT-FA rate.

4.2. Analysis with RM-3000 dataset

From the experiments in Section 4.1, loss functions CE and CHF are selected for further analysis in the larger vocabulary tests which use the RM-**3000** dataset. Experiments are performed in babble noise at SNRs of -5 dB, 0 dB and +5 dB, with LC set to 5 dB lower than the select SNR. Table 4 shows objective performance across all system configurations.

Table 4: *Classification accuracy (in %), HIT-FA (in %) and ESTOI scores for the RM-*3000 *dataset in babble noise at* -5 dB*,* 0 dB *and* +5 dB.

SNR	Feat	Loss	Acc	HIT-FA (FA)	ESTOI
-5 dB	A	CE	**90.3**	71.2 (4.8)	**46.9**
		CHF	88.8	**76.2** (10.6)	46.5
	V	CE	**84.7**	58.5 (9.5)	38.8
		CHF	82.6	**65.7** (17.8)	**40.0**
	AV	CE	**90.7**	73.6 (5.3)	**50.5**
		CHF	89.1	**78.0** (10.9)	**50.5**
	unprocessed audio				22.0
0 dB	A	CE	**91.7**	76.2 (4.6)	59.6
		CHF	90.5	**80.4** (9.2)	**60.5**
	V	CE	**84.7**	58.2 (9.4)	44.5
		CHF	82.8	**65.5** (17.2)	**48.0**
	AV	CE	**91.8**	77.1 (4.7)	61.3
		CHF	90.6	**81.2** (9.6)	**62.2**
	unprocessed audio				35.4
+5 dB	A	CE	**92.4**	78.9 (4.5)	68.8
		CHF	91.3	**82.7** (8.8)	**70.8**
	V	CE	**84.7**	58.5 (9.5)	50.1
		CHF	82.6	**65.7** (17.6)	**56.2**
	AV	CE	**92.4**	78.8 (4.5)	69.1
		CHF	91.3	**82.9** (9.0)	**71.5**
	unprocessed audio				50.7

As with the experiments with GRID (Section 4.1) supplementing audio with visual information provides best performance across all measures for all SNRs, confirming that combining audio and visual features provides a robust complimentary feature set. Largest gains were found at low SNRs, at -5 dB a gain of 3.6 in ESTOI was achieved over audio-only, providing an overall gain of 28.5 over unprocessed. The performance benefit of audio-visual over audio-only is less using the RM-3000 dataset compared to the GRID dataset, due to the overall decrease in performance of the visual features shown through visual-only experiments. This is due to the larger variability associated with the RM-3000 dataset compared to the GRID dataset making it more challenging for the DNN to distinguish the similar mouth shapes associated within the visual feature.

Similar to GRID, large gains in intelligibility over the unprocessed audio were found with both loss functions. When the SNR is very low, the CE loss function is best, and at all other SNRs the hybrid CHF loss function outperforms CE.

5. Conclusions

This work has examined the effect on intelligibility of including visual information in binary mask estimation for speech enhancement. It was found that all systems provide large gains in intelligibility over unprocessed, with largest gains found at lower SNRs. Combining both audio and visual modalities into a single bimodal audio-visual system provides largest gains across all noise types, SNRs and datasets, confirming that combining audio and visual features provides a robust complimentary feature set.

This work has also proposed two new perceptually motivated loss function for DNN-based mask estimation inspired by the HIT-FA rate which is known to correlate closely to speech intelligibility. A hybrid cross-entropy HIT-FA loss function (CHF) was proposed to reduce the bias found within binary cross-entropy by adjusting the ratio between 1s and 0s inspired by HIT-FA. Application of the proposed loss functions was evaluated on a small vocabulary (GRID) and large vocabulary (RM-3000) dataset. Evaluations using classification accuracy, HIT-FA rate and ESTOI reveal that the proposed loss functions provide performance gains in HIT-FA and ESTOI over the standard binary cross-entropy loss function, with peak performance found with our hybrid loss function (CHF) across both datasets.

6. Acknowledgements

We wish to thank the UK Home Office – Centre for Applied Science and Technology, for supporting this work. The research presented in this paper was carried out on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at the University of East Anglia.

7. References

[1] J. Barker, L. Josifovski, M. Cooke, and P. D. Green, "Soft decisions in missing data techniques for robust automatic speech recognition." in *INTERSPEECH*, 2000, pp. 373–376.

[2] J. Barker, M. Cooke, and P. D. Green, "Robust ASR based on clean speech models: an evaluation of missing data techniques for connected digit recognition in noise." in *INTERSPEECH*, 2001, pp. 213–217.

[3] B. C. Moore, *Cochlear hearing loss: physiological, psychological and technical issues.* John Wiley & Sons, 2007.

[4] P. C. Loizou, *Speech enhancement: theory and practice.* CRC press, 2013.

[5] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms." *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, 2007.

[6] H. Levitt, "Noise reduction in hearing aids: a review." *Journal of rehabilitation research and development*, vol. 38, no. 1, p. 111, 2001.

[7] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications.* Wiley-IEEE Press, 2006.

[8] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound.* MIT press, 1994.

[9] M. Ahmadi, V. L. Gross, and D. G. Sinex, "Perceptual learning for speech in noise after application of binary time-frequency masks." *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1687–1692, 2013.

[10] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation." *The Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 4007–4018, 2006.

[11] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction." *The Journal of the Acoustical Society of America*, vol. 123, no. 3, pp. 1673–1682, 2008.

[12] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking." *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2336–2347, 2009.

[13] K. Han and D. Wang, "A classification based approach to speech segregation." *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3475–3483, 2012.

[14] K. Thangthai, R. W. Harvey, S. J. Cox, and B.-J. Theobald, "Improving lip-reading performance for robust audiovisual speech recognition using dnns." in *FAAVSP*, 2015, pp. 127–131.

[15] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audio-visual speech," in *Proceedings of the IEEE*, vol. 91, no. 9, 2003, pp. 1306–1326.

[16] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 1, pp. 1260–1273, 2002.

[17] R. Y. Rubinstein and D. P. Kroese, *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning.* Springer Science & Business Media, 2013.

[18] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners." *The Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 3029–3038, 2013.

[19] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners." *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[20] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios." *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 12, pp. 1993–2002, 2014.

[21] E. W. Healy, S. E. Yoho, J. Chen, Y. Wang, and D. Wang, "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type." *The Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. 1660–1669, 2015.

[22] J. Chen, Y. Wang, and D. Wang, "Noise perturbation for supervised speech separation." *Speech Communication*, vol. 78, pp. 1–10, 2016.

[23] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function." in *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, 1987.

[24] Y. Lan, R. Harvey, B. Theobald, E.-J. Ong, and R. Bowden, "Comparing visual features for lipreading," in *International Conference on Auditory-Visual Speech Processing 2009*, 2009, pp. 102–106.

[25] D. Websdale and B. Milner, "Analysing the importance of different visual feature coefficients," in *FAAVSP*, 2015.

[26] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[27] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing.*, vol. 34, no. 1, pp. 52–59, Feb 1986.

[28] B. Hanson and T. Applebaum, "Robust speaker-independent word recognition using static, dynamic and acceleration features: Experiments with Lombard and noisy speech." in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. IEEE, 1990, pp. 857–860.

[29] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation." *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 7, pp. 1381–1390, 2013.

[30] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition." *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[31] D. Howell, S. Cox, and B. Theobald, "Visual units and confusion modelling for automatic lip-reading," *Image and Vision Computing*, vol. 51, pp. 1–12, 2016.

[32] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, "The darpa 1000-word resource management database for continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*. IEEE, 1988, pp. 651–654.

[33] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, no. 1, 2013.

[34] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[35] L. Prechelt, "Early stopping-but when?" in *Neural Networks: Tricks of the trade*. Springer, 1998, pp. 55–69.

[36] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, 2012.

[37] S. Dieleman, J. Schlter, C. Raffel, E. Olson, S. K. Snderby, D. Nouri *et al.*, "Lasagne: First release." Aug. 2015. [Online]. Available: http://dx.doi.org/10.5281/zenodo.27878

[38] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016. [Online]. Available: http://arxiv.org/abs/1605.02688

[39] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.