Fast Lip Feature Extraction Using Psychologically Motivated Gabor Features

Andrew Abel
Computing Science and Software Engineering
Xi'an Jiaotong-Liverpool University
Suzhou, China
andrew.abel@xjtlu.edu.cn

Chengxiang Gao
Computing Science and Software Engineering
Xi'an Jiaotong-Liverpool University
Suzhou, China
chengxiang.gao16@student.xjtlu.edu.cn

Leslie Smith
Faculty of Natural Sciences
University of Stirling
Stirling, Scotland
lss@cs.stir.ac.uk

Roger Watt
Faculty of Natural Sciences
University of Stirling
Stirling, Scotland
r.j.watt@stir.ac.uk

Amir Hussain
Faculty of Natural Sciences
University of Stirling
Stirling, Scotland
ahu@cs.stir.ac.uk

***Abstract*—The extraction of relevant lip features is of continuing interest in the speech domain. Using end-to-end feature extraction can produce good results, but at the cost of the results being difficult for humans to comprehend and relate to. We present a new, lightweight feature extraction approach, motivated by glimpse based psychological research into facial barcodes. This allows for 3D geometric features to be produced using Gabor based image patches. This new approach can successfully extract lip features with a minimum of processing, with parameters that can be quickly adapted and used for detailed analysis, and with preliminary results showing successful feature extraction from a range of different speakers. These features can be generated online without the need for trained models, and are also robust and can recover from errors, making them suitable for real world speech analysis.**

***Index Terms*—image processing, barcodes, gabor, lip-reading, word features**

I. Introduction

Lip features are an important part of speech communication [1]. One particular topic of interest is the use of lip features for speech processing [2], [3], [4], [5]. A variety of methods have been developed, including 2D-DCT [6], which was found to correlate highly to audio vectors [7], [3], optical flow [8], [9], [10], shape and/or appearance models [11], and active contour models [12]. Active Appearance Models [11] require significant offline training, some approaches [13] are capable of online training, but sometimes lack precision. Recent research has successfully used shape models as part of an overall visual feature vector [14], and adaptive online trackers can track a region of interest (ROI) [15].

Geometric features (the width and height of the mouth [5]) can be calculated, but require the mouth to be correctly identified, which uses various shape and appearance models [11], [13], or markers [5]. However, geometric features are generally 2-dimensional, and detection requires other modelling techniques.

In recent years, one increasingly popular approach has been to use Convolutional Neural Networks (CNN) for end-to-end lipreading. These use the image directly [16], but do not allow us to establish direct relationships explaining how words are created, and does not output features in a useful form for analysis. Another approach is to use a trained CNN to extract features. This is done by applying the trained CNN to the image and then using the weights of the last layer as an input into another network, such as an LSTM [17]. Another approach that has been used is LipNet [18], which is a trained CNN that has produced very good results, reporting 93.4% accuracy on sentences from the Grid corpus. However, other research has attempted to use this work, with limited results [19], possibly due to the non-explainable nature of the LipNet features..

It should be noted that some of these CNN based approaches deliver very good results, and increasingly, the trend, as seen in other fields is to use very large datasets and trained CNNs to produce good results, as shown by LipNet producing 93.4% on sentences from the Grid corpus, using only the lip features, far beyond human lipreading performance. However, there are problems with this approach. Using end-to-end learning with CNNs means that they learn their own features rather than using features directly. This means that the results, while good, can be very difficult to explain and justify. Although the researchers behind LipNet apply saliency maps to identify regions of the mouth focused on by the system, this still makes it difficult to explain the results. This is also a problem with other approaches such as DCT. The features do not easily map to human perception, and are not easy to explain.

This is important for the concept of explainable AI (XAI) [20], where as well as producing the results, we also wish to be able to justify our decision. It is also important for psychological and linguistic speech research, where being able

Andrew Abel performed part of this work while at the University of Stirling. The work in this paper was partially funded by EPSRC Grant EP/M026981/1 (AV-COGHEAR).

to explain and map the results, as well as gain insights, is as important as being able to generate excellent results. This means there is a requirement for visual features which are fast to generate, do not need complex training, are lightweight, and can be both used for machine learning, and can also be interpreted by human experts.

In this paper, we therefore propose a fast, lightweight approach for generating three-dimensional lip features, which is less reliant on accurate mouth modelling, and creates features in a format that can be analysed. We use psychologically motivated Gabor filtering that can identify a range of useful features for further analysis. Gabor features have been used previously for lip feature work, notably by Hursig et al. [21], who used Gabor features for identifying the lip region correctly. The key difference is that we aim to obtain speech vectors, rather than region identification. Feature extraction is quick and robust and has been successfully applied to a variety of faces. It can provide simple and understandable mouth movement information, including mouth opening area, and depth. This new approach has many potential applications, including speech recognition, synthesis, tracking, and linguistics. For example, the simple and quickly calculated outputs of this system could be used to determine changes in speech effort, or identify different accents (by identifying unique speech features that are easily visually distinguishable, such as saying the same word in different ways). They could also be used to inform mouth movements with regard to the generation of synthetic talking heads. We demonstrate the potential of our approach with word analysis using our extracted features, and represent a potential new feature-set to be used with machine learning.

II. Psychologically Motivated Gabor Features

Humans can recognise faces using distinctive facial features. Independent perceptual attributes of faces can be encoded using the concept of face space [22] [23], where distinctiveness is encoded as the difference from an overall average. The biological approach to face recognition provides evidence that humans use early-stage image processing, such as edges and lines [24]. Dakin and Watt [23] examined this with Gabor filters. They used different filter orientations, and found that horizontal features were the most informative, that distinct facial features could be robustly detected, and that vertical slices of the centre of the face could form a distinctive "barcode". This was developed further by [25].

Based on the motivation behind these barcode based features, the coarse distinctions between facial features can also be similarly applied to much finer detailed features [23]. There are clear differences between the features, such as the lips, teeth, philtrum, and mentolabial sulcus. The contrasts present in different mouth openings allows for quick and accurate mouth feature information to be obtained, with a three dimensional representation of the mouth opening possible. Therefore, the principle of horizontal Gabor features can also be applied to lip specific feature extraction.

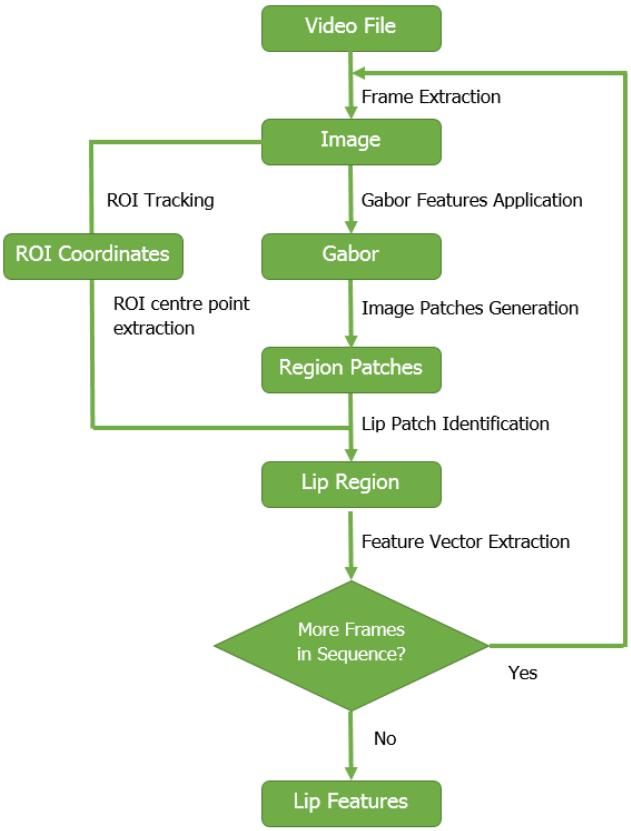

Fig. 1. Key stages of lip feature extraction.

III. Proposed Approach

A. ROI Identification and Tracking

The key steps of our approach are shown in Fig. 1. Given a sequence of images $I_n(n = 1 \dots N)$ extracted from a video file, the lip region must be tracked. As ROI identification is not a key contribution of this paper, we follow previous research [2], [26] and use a Viola-Jones detector and an online shape model [15], similar to previous Gabor feature lip research [21]. This outputs a coarse 2-dimensional lip region for each image frame, represented as the four x and y coordinate pairs ($L_x^n(1, 2, 3, 4)$ and $L_y^n(1, 2, 3, 4)$ respectively). From these, we can identify C_L^n, the ROI centre point for each frame.

B. Gabor Feature Generation

Similar to Dakin and Watt [23], we calculate horizontal Gabor features, using a Fast Fourier Transform. This generates positive and negative going real and imaginary components, and here we use the real component. Each image is converted to greyscale, and Gabor filtering is applied, see Fig. 2 (b). To reduce small values such as background noise and regulate the size of the image patches, threshold is applied to the initial transform, as shown in Fig. 2 (c). Several parameters are required. These tend to only need to be adjusted when a different corpus is used, if for example, video frames are very

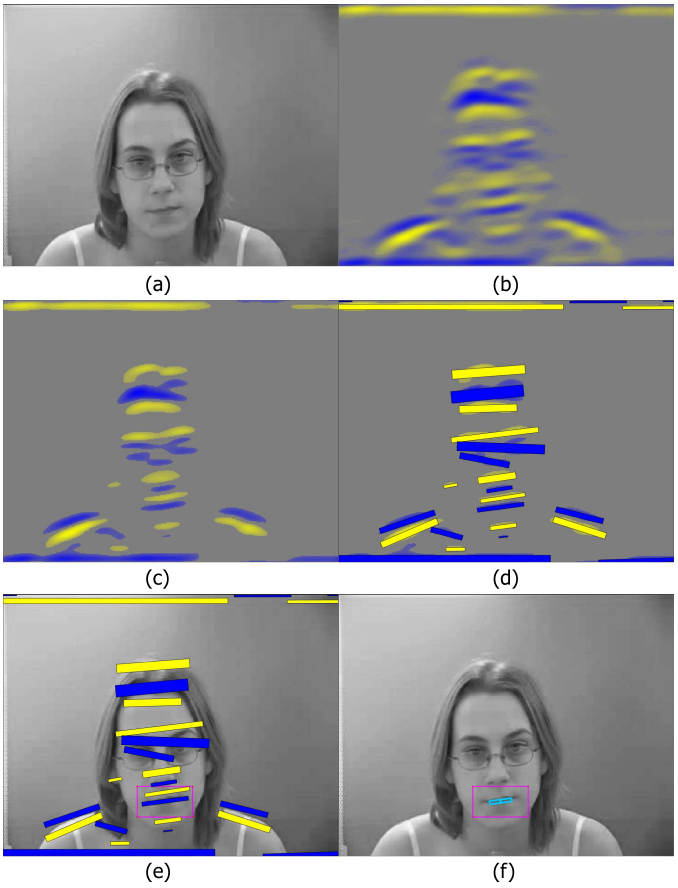

Fig. 2. The lip patch generation process, showing (a) the original greyscale image, prior to processing, (b) the real component of the Gabor features, (c) the thresholded image, (d) the resulting Gabor image patches, (e) the image patches and the tracked ROI box for that frame, and (f) the final chosen lip patch.

differently sized, or a speaker is sitting at a prominent angle, or at a different distance from the camera:

Wavelength λ - The individual Gabor wavelength. This can feasibly be between 2 and 20+. The exact parameter depends on image size.

Threshold t - The filtering threshold, to ignore minor face features and background noise. The range is 0 to 1, and a value between 0.05 and 0.3 has been found to be effective.

Orientation Θ - The face angle in degrees. If the speaker has their head straight, then 0 (horizontal) is suitable, but a slight angle, commonly $\Theta = 5$, may be needed.

Min. Patch Area P_{MIN} - The minimum size needed for a region patch. This can be useful when speakers have (for example) a prominent chin or teeth. Generally, a value of around 50 to 100 is suitable.

C. Image Region Patches and Relevant Patch Identification

After filtering and thresholding, the most prominent regions (i.e. the local extrema in the filter outputs) are calculated, and these regions are grouped and represented by rectangular image patches. These are calculated using the filtered real component of the transformed image, shown in Fig. 2 (c).

Given a filtered image, patches are then created from these components. Firstly, using the Matlab "bwconncomp" function, with 8 degrees of connectivity, and the "regionprops" function, creates R groups of connected pixels. The result is a matrix of pixel locations, Q^X and Q^Y, and values Q^V for each grouping, G_r. For each G_r, the area A_r is defined as the number of pixels in each G_r, so A_r. The mass is calculated using each pixel value as,

$$M_r = \sum_{p=1}^{P} Q_p^V \tag{1}$$

The centre coordinates of each patch, X_r and Y_r are calculated using both pixel coordinates (p_x, p_y) and pixel values. As some of the pixels at the edges of the regions may not be as strongly connected, then this is taken into account,

$$X_r = \sum_{p=1}^{A_r} (p_x * Q_p^V)/M_r \tag{2}$$

$$Y_r = \sum_{p=1}^{A_r} (p_y * Q_p^V)/M_r \tag{3}$$

The variance is calculated, $\sigma^2_{X_r}, \sigma^2_{Y_r}$, as is the covariance, which is given in equation 4,

$$\sigma(X_r, Y_r) = \sum_{p=1}^{A_r} (p_x * p_y * Q_p^V)/M_r - X_r * Y_r \tag{4}$$

The patches are not always horizontal or vertical, and they have an orientation, Θ, calculated using covariance and variance,

$$\Theta = tan^{-1}\left(2 * \sigma(X_r, Y_r), (\sigma^2_{X_r} - \sigma^2_{Y_r})\right)/2 \tag{5}$$

Θ can then be used to calculate width and height of each patch. The width is calculated as,

$$W_r = \sqrt{(abs(w_r) + 0.5/\pi)} \tag{6}$$

where w_r is defined as,

$$\begin{aligned} w_r = (X_r^2 - (X_r)^2) * \cos^2\Theta \\ +2 * \sigma(X_r, Y_r). * (\cos\Theta * \sin\Theta) \\ +\sigma^2_{Y_r} * (\sin^2\Theta)) \end{aligned} \tag{7}$$

This also requires the squared value,

$$X_r^2 = \sum_{p=1}^{A_r} (p_x^2 Q_p^V)/M_r \tag{8}$$

The height, H_r is calculated with a similar process,

$$H_r = \sqrt{(abs(h_r) + 0.5/\pi)} \tag{9}$$

$$h_r = (Y_r^2 - (Y_r)^2) * \cos^2\Theta$$
$$-2 * \sigma(X_r, Y_r). * (\cos\Theta * \sin\Theta) \quad (10)$$
$$+\sigma_{X_r}^2 * (\sin^2\Theta))$$

$$Y_r^2 = \sum_{p=1}^{A_r} (p_y^2 Q_p^V)/M_r \quad (11)$$

These properties allow for the creation of patches. An example is shown in Fig. 2 (d), showing all the resulting patches generate from this image. It can be seen that there are patches generated around the hair, eyes, nose, mouth and shoulders. Of key interest in this paper is the patch corresponding to the mouth opening.

For each patch, a number of values are generated. These are quick to generate, and can be used for analysis. For this work, the most relevant are:

Width The width of the lip region, W_r

Area The area is A_r. The height is constrained by λ, and tends to only show big changes, but can be a good measure of mouth opening.

Mass This is related to intensity, and is defined as M_r. It effectively shows the mouth depth, providing 3D representation. It can distinguish between a closed mouth, an open mouth showing teeth (as in an 'ee' sound), and an open mouth making an 'oh' or 'ah' sound.

Xpos The x position, X_r identifies the mean x-position of the pixels in the patch, i.e. the centre position of the x-co-ordinate. This can be useful, along with the y-position, for tracking speaker movement.

Ypos The y position, Y_r identifies the mean y-position of the pixels in the patch, i.e. the centre position of the y-co-ordinate. This can be very useful for tracking speech. For example, with tonal languages, research has identified [27] that during certain tones, speakers dip their heads.

Θ As discussed previously, Θ is used to calculate the orientation of each patch. This is different from the orientation of the Gabor wave Θ. Here, Θ corresponds to each patch orientation, so for example, each shoulder in Fig. 2 (d) would have a different orientation.

As discussed, to calculate the ROI centre point, $C_L^n(x, y)$, is calculated. This can be used to identify the lip region patch. To do this, for each frame, the ROI centre point, $C_L^n(x, y)$ is compared to each r-th object of X and Y for each n-th frame to identify the closest patch. This patch is then used for further analysis. This is shown in Fig. 2 (f), showing the ROI as a pink rectangle, and then the chosen patch (the mouth opening) in blue. The complete process is shown in Fig. 2.

D. Extracted Features

The output can be visualised as a sequence of frames, showing the ROI and the lip features, as discussed above.

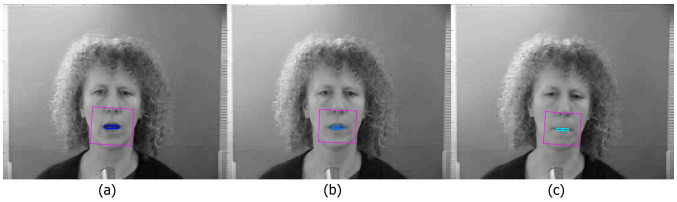

Fig. 3. (a) Example of a wide open mouth, with a large box and dark colour, (b) an open mouth, but without depth (due to the teeth being visible), reflected by the lighter colour, (c) a closed mouth.

Fig. 3 shows an example from the Grid Corpus [28]. It is notable, that as well as being able to quickly calculate the mouth region, the mass can also be calculated, showing how open the mouth is. To visualise this, the mass is used to adjust the colour of the lip region patch, with a lighter blue being used for a closed mouth, and an darker blue being used for a closed mouth. Fig. 3 (a) shows a wide open mouth, with a large box and dark colour, Fig. 3 (b) shows an open mouth, but without depth (due to the teeth being visible), reflected by the lighter colour. Finally Fig. 3 (c) shows a closed mouth. This is reflected in the reduced size of the patch and light colour, showing a quick and simple 3-dimensional representation of the mouth features.

These outputs can also be visualised as vectors, as shown in Fig. 4 (a), and (b), which show the change in width in Fig. 4 (a), and the change in mass in figure 4 (b). It can be seen that in each frame, a single value is generated for each feature, and that there is a natural flow over the course of a speech sentence.

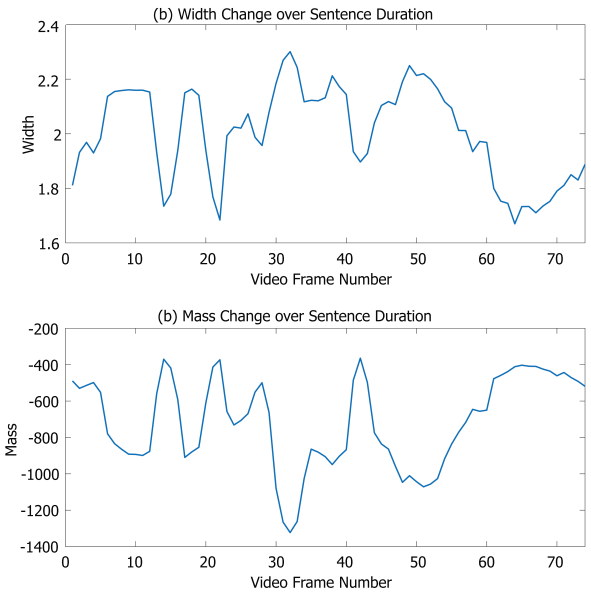

Fig. 4. Examples of (a) width change, and (b) mass change, over a single speech sentence.

In Fig. 4 the x-axes corresponds to the number of frames,

with one data point for each frame, with the y-axes representing amplitude. The amplitude changes are of interest, as they show the differences between individual frames, and also between sentences and speakers, which is of interest for further analysis. Here, the width shows that the mouth gets narrower and wider depending on the word, and the mass also shows the opening and closing of the mouth. This shows that we are able to analyse speech in more detail, in a way that features such as CNN or DCT features are unable to do.

IV. Results and Discussion

This section focusses on a visual examination of results. The aim is not to run a comparison with AAM, DCT, or CNN features, because these features can not easily be used for visual analysis. This paper presents the initial features, and considers their relevance for speech analysis. Future work will investigate these features for machine learning, when comparisons can be more easily made to other approaches.

A. Tracking and Parameter Selection

We successfully tracked many (150+) videos from multiple corpora, including Grid [28], and VidTIMIT [29], which were chosen due to their wide use in speech processing research. The tracker was found to be effective for our needs, although could easily be replaced by other approaches. As discussed, we used the tracker used in previous research, [2], [26], which was a Viola-Jones detector, with a shape tracker. There are known limitations with this approach, and future research will investigate alternative methods.

In this paper, we chose to present the initial features, and also perform some word specific analysis, as The GRID corpus includes alignment files, which enables word specific analysis. Due to space limitations, we cannot show detailed results, but parameters were kept as consistent as possible, with only slight adjustments. In almost all cases, the Gabor wavelength λ was set to 5, with a slightly larger λ for higher resolution frames, and in almost all cases, the patch area was set to 50. The threshold t varied between 0.14 and 0.25 depending on experimentation, and Θ was generally set to 0, although setting it to 5 is useful when the speaker is at a slight angle. As an example, we present some specific parameters from two established corpora in Table I.

TABLE I
Summary sentences used for testing, with parameters used.

Corpus	Speaker	Sent. ID	λ	t	Θ	P_{MIN}
Grid	s1	bbaf2n	5	0.14	0	50
		bbaf3s	5	0.14	0	50
		swwv8p	5	0.14	0	50
		swwv9a	5	0.14	0	50
	s2	bbaf1n	5	0.14	0	50
		bbaf2s	5	0.14	0	50
		swwv7p	5	0.14	0	50
		swwv8a	5	0.14	0	50
	s6	bbae7n	5	0.14	0	50
		bbae8s	5	0.14	0	50
		swwv3p	5	0.25	5	50
		swwv4a	5	0.14	0	50
	s7	bbae6n	5	0.25	5	50
		bbae7s	5	0.25	5	50
		swwv3a	5	0.25	5	50
		swwvzn	5	0.25	0	50
	s15	bbad8n	5	0.14	0	50
		bbad9s	5	0.14	5	50
		swwu4p	5	0.14	0	50
		swwu5a	5	0.14	0	50
	s26	bbac7n	5	0.10	0	50
		bbac8s	5	0.10	0	50
		swwt3p	5	0.10	5	50
		swwt4a	5	0.10	0	50
VidTIMIT	mrjo1	sa1	5	0.14	0	50
	fadg0	sa1	5	0.20	0	50

Table I shows the parameters used for the different corpora. We can see that for the majority of cases, the parameters are fairly similar. The majority of Grid corpus videos use the same parameters, with some variations for speaker s6, where one video is further away from the camera than the others, which required a change in the threshold. In speaker s7, the speaker speaks with their mouth at an angle, as shown in figure 6. This required a change in threshold, and also a slight change in orientation to optimise results. Finally, speaker s26 was a special case due to the specific combination of skin colour, lip features, and facial hair (a prominent moustache). This meant that the parameters were initially difficult to identify successfully, and in 3 of the 4 videos, the system needed to be customised to allow for the threshold parameter to be changed during the course of the video, which did not need to be done for any other speaker. Although adequate results could eventually be achieved, there were some limitations with this work that require further investigation.

For the VidTIMIT corpus, some parameters needed to be changed for speaker fadg0, as she enunciates clearly, and so her mouth region was more prominent than other videos, requiring a change in threshold. Otherwise, other videos used standard parameters. Finally, a number of custom videos were recorded to test the system in unpredictable environments, as the existing Grid and VidTIMIT corpora tended to be in optimal environments. It was found in these that there were some issues identified. For example, if the speaker was close to the camera, then the wavelength needed to be increased. Other speakers had particularly prominent teeth when they were close to the camera, and here, the minimum patch area could be adjusted. In addition, the same limitation was present with this feature extraction approach as for other approaches in the literature. Fig. 5 shows two examples of this. The first speaker has a moustache covering his mouth, making tracking impossible. The second speaker has hair covering his face, is looking down, rather than at the camera, and has a lot of facial hair, also making feature extraction challenging.

However, this is an issue common to many other feature extraction approaches, and as our approach is not reliant on a model, recovery from a glitch is easy. So if an incorrect patch is identified in one frame, or a patch is not identified, this error is not compounded in subsequent frames, as each frame is calculated individually. As an example of successful tracking, Fig. 6 shows a number of frames from sentence bbae6n by Grid speaker s7. Of most interest is that in addition to the

Fig. 5. Two examples of speakers recorded for trials that could not be easily used for feature extraction.

conventional 2D information (lip widths, area), the colour change shows when the mouth is wide open (bottom left), when the mouth is totally closed (top left and bottom right), and various transitional phases, which are clearly represented by the colour change, with the lip tracker colour being lighter for more closed mouths, and darker for more open mouths. This is an important additional feature that helps to build a 3D representation of the mouth region.

Fig. 6. Tracking with Grid speaker S7 for a single sentence, showing change in mouth opening size, and also the slight angle required due to the speaker having their head at an angle.

Overall, we found that our approach could accurately track lip features over a range of different speakers, genders, sentences, and corpora. The parameters are flexible and mean that it is straightforward to adjust for individual videos. In this paper, we therefore focus on demonstrating preliminary example results of the tracking process. In all cases, the features were all extracted from the video file without any offline training being required, although some videos had their parameters adjusted and were re-run. These results are intended to show that speech data can be analysed by the human eye, and that our approach can produce consistent results.

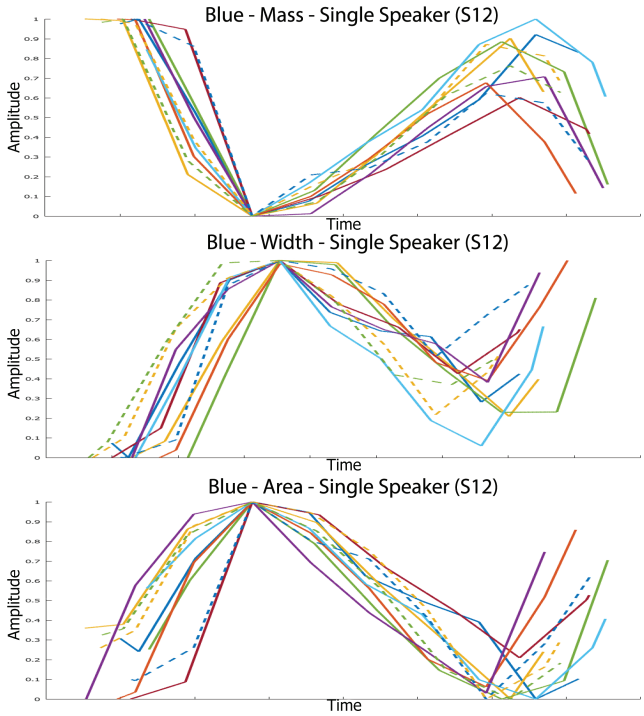

Fig. 7. 10 normalised sentences from speaker 12, showing from top, mass, width, and area for the word 'blue'.

B. *Individual Word Analysis*

We aim to produce simple data that can demonstrate word relationships and then be used for future feature extraction. As discussed previously, we extracted a number of sentences from the Grid corpus. As speakers have different speech rates and mouth sizes, we normalise over time and amplitude, and use word alignment data (provided by the Grid corpus) to identify individual words. Obviously, as these were manually identified by the authors of the Grid corpus, this means that the words do not start at the exact same time (for example, some words may start immediately, where others may have a small pause). As the alignment data is not fully precise, we adjust the x-axes slightly where appropriate to match peaks. It should be noted that this type of analysis is hard to compare to other features, such as DCT or CNN features, as these features can not be simply and clearly visualised.

To demonstrate these results, we have chosen some example words. The first 3 plots in Fig. 7 show 10 sentences from Grid speaker 12, showing the mass, width, and area of the word 'blue'. Fig. 7 (top) shows the mass (the 3D feature). The mouth closes on the 'b' plosive which is shown in the minima, and then gradually opens. Fig. 7 (2nd) and (3rd) show the width and area changes. They are very similar, but the width has a smaller rate of change, with increased width during the closed mouth of 'b', and a slight reduction over time. Of great interest is the consistent pattern that can be seen for all 10 sentences.

In Fig. 7, we used speaker 12 as a representative example. Other speakers were also found to exhibit very similar patterns. We demonstrate this in a single word with limited changes.

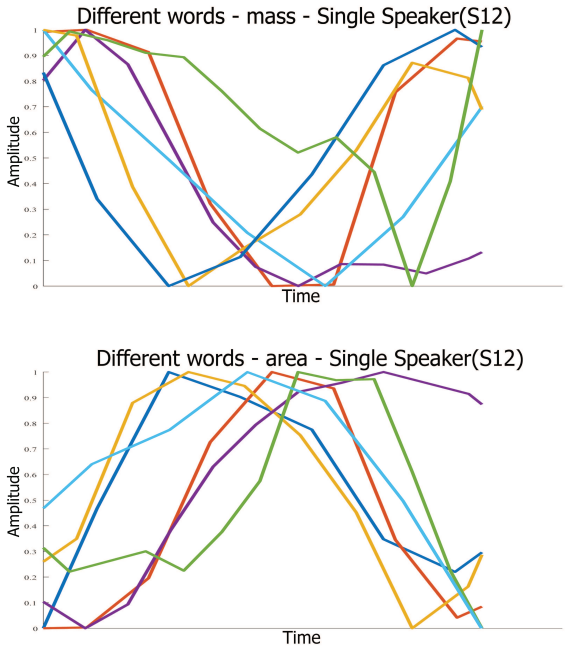

Fig. 8. Different words from the same speaker, showing mass (top) and area (bottom) for 6 different words.

Fig. 9 shows the word 'eight' said by 6 different speakers. It can be seen that there are clear patterns, with one exception, in a single sentence, where the word was found to be a little different, possibly due to pre-voicing the next word. The word 'please' in Fig. 7 (4th) is similar. The mass peaks when the mouth opens during pre-voicing, followed by a closing for the plosive of 'p'. The area and width are again very similar, with only the area being shown here, with the narrowest area before the 'p' is formed, which expands around the 'ee' stage, before closing slightly for the 's' part. Again, all 10 sentences here show a similar pattern, showing a pattern for all 10 speakers.

Finally, we plotted 6 different words from the same speaker in Fig. 8 , showing 'at', 'bin', 'blue', 'now', 'nine', and 'q'. As we can see, despite the normalisation, there is no clear pattern, unlike the other examples above, showing that our approach can identify individual words very simply.

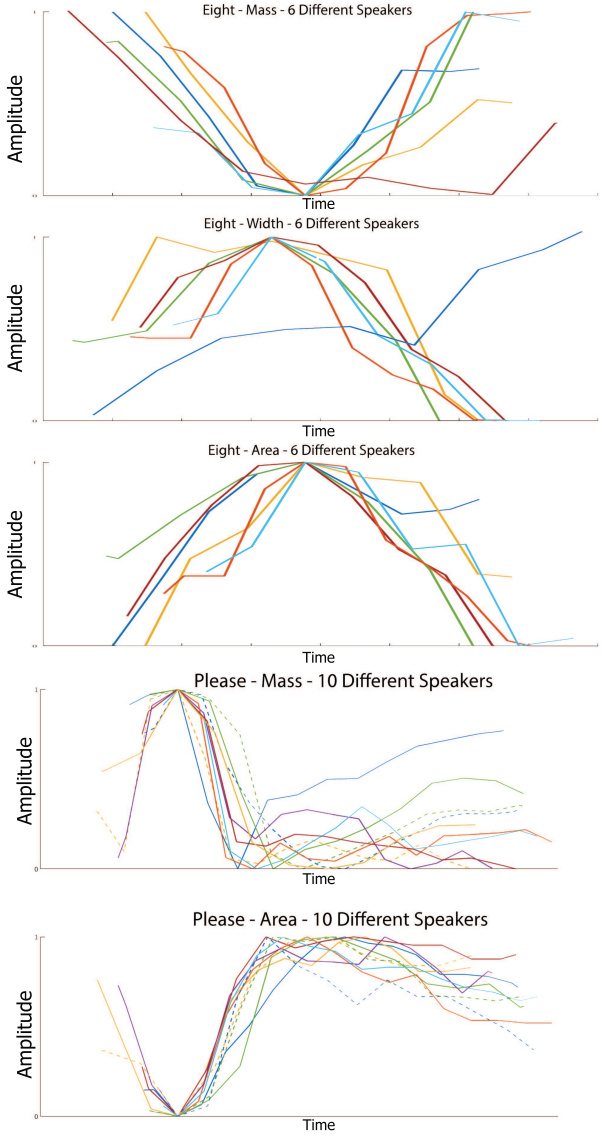

Fig. 9. 6 normalised sentences from Grid speakers 15, 16, 17, 18, 19, 23, showing (from top) mass, width, and area for the word 'eight'. We also show examples from 10 different speakers for the word 'please', showing mass and area.

V. Conclusions and Future Directions

We presented a very lightweight and quick approach to generating 3 dimensional lip features. Example results showed that these features can represent words in a way that can be distinctly and consistently visualised, and can be applied to a wide number of different speakers. However, as with many similar approaches, there are limitations, although due to space limits, we were unable to discuss these in detail. We found that speakers with facial hair could cause problems, as did head turning and looking away. However, as there is no model used, the tracking can easily recover from short term errors.

The results presented in this paper focused on speech analysis and how to use the speech features to carry out a manual analysis. This is the initial presentation of these features, and the key difference is that using the human eye, key differences between words can be identified, which are consistent across the same speaker saying the same word several times, and also across different speakers. The identification can be carried out visually, and the features can be explained to non-experts. The features are also quick to calculate and lightweight.

The next step with these features is to experiment with deeper machine learning. The aim is to run deeper comparisons for speech processing, considering word recognition, frame based speech estimation [26], or speech filtering [30]. This will involve using LSTM (Long Short Term Memory) based machine learning, and running comparisons with other

approaches. The key benefit of this approach is the simplicity of calculation, and the explainable nature of the features. This work has a number of practical applications, such as linguistic (pronunciation) training and further speech analysis. Further future work will further investigate these features, carry out speech recognition tests, and improve the system to prevent glitches.

REFERENCES

[1] W. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[2] A. Abel and A. Hussain, *Cognitively inspired audiovisual speech filtering: towards an intelligent, fuzzy based, multimodal, two-stage speech enhancement system*, 1st ed. SpringerBriefs in Cognitive Computation, Springer International Publishing, 2015, vol. 5.

[3] I. Almajai and B. Milner, "Effective visually-derived wiener filtering for audio-visual speech processing." in *AVSP*, 2009, pp. 134–139.

[4] G. Potamianos, C. Neti, and S. Deligne, "Joint Audio-Visual Speech Processing for Recognition and Enhancement," in *AVSP 2003-International Conference on Auditory-Visual Speech Processing*, 2003, pp. 95–104.

[5] L. Girin, J. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *The Journal of the Acoustical Society of America*, vol. 109, p. 3007, 2001.

[6] N. Ahmed, T. Natarajan, and K. Rao, "Discrete cosine transfom," *Computers, IEEE Transactions on*, vol. 100, no. 1, pp. 90–93, 1974.

[7] A. Abel and A. Hussain, "Novel two-stage audiovisual speech filtering in noisy environments," *Cognitive Computation*, pp. 1–18, 2013.

[8] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[9] B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision." in *IJCAI*, vol. 81, 1981, pp. 674–679.

[10] M. Sargın, E. Erzin, Y. Yemez, and A. Tekalp, "Lip feature extraction based on audio-visual correlation," in *Proc. EUSIPCO*, vol. 2005, 2005.

[11] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *Computer Vision–ECCV 98*, pp. 484–498, 1998.

[12] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[13] Q. Nguyen and M. Milgram, "Semi adaptive appearance models for lip tracking," in *ICIP09*, 2009, pp. 2437–2440.

[14] B. Milner and D. Websdale, "Analysing the importance of different visual feature coefficients," *FAAVSP-The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing*, 2015.

[15] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125–141, 2008.

[16] Y. Li, Y. Takashima, T. Takiguchi, and Y. Ariki, "Lip reading using a dynamic feature of lip images and convolutional neural networks," in *Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on*. IEEE, 2016, pp. 1–6.

[17] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," *arXiv preprint arXiv:1611.05358*, 2016.

[18] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: End-to-end sentence-level lipreading," *arXiv:1611.01599*, 2016.

[19] M. Faisal and S. Manzoor, "Deep learning for lip reading using audio-visual information for urdu language," *arXiv preprint arXiv:1802.05521*, 2018.

[20] T. Miller, P. Howe, and L. Sonenberg, "Explainable AI: Beware of Inmates Running the Asylum," in *IJCAI-17 Workshop on Explainable AI (XAI)*, 2017.

[21] C. Hursig, Robert E and Zhang, Jane Xiaozheng and Kam, "Lip Localization Algorithm Using Gabor Filters," in *International Conference on Image Processing and Computer Vision*, 2011, pp. 357–362.

[22] D. A. Leopold, A. J. O'Toole, T. Vetter, and V. Blanz, "Prototype-referenced shape encoding revealed by high-level aftereffects," *Nature neuroscience*, vol. 4, no. 1, p. 89, 2001.

[23] S. C. Dakin and R. J. Watt, "Biological bar codes in human faces," *Journal of Vision*, vol. 9, no. 4, pp. 2–2, 2009.

[24] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of physiology*, vol. 195, no. 1, pp. 215–243, 1968.

[25] Y. Matveev, G. Kukharev, N. Shchegoleva, and S.-P. Electrotechnical, "A simple method for generating facial barcodes," in *22nd Intern. Conf. in Central Europe on Computer Graphics, Visualization and Computer Vision, WSCG*, 2014, pp. 213–220.

[26] A. Abel, R. Marxer, A. Hussain, J. Barker, R. Watt, B. Whitmer, P. Derleth, and A. Hussain, "A Data Driven Approach to Audiovisual Speech Mapping," in *Advances in Brain Inspired Cognitive Systems: 8th International Conference, BICS 2016, Beijing, China, November 28-30, 2016, Proceedings 8*. Springer International Publishing, 2016, pp. 331–342.

[27] T. H. Chen and D. W. Massaro, "Seeing pitch: Visual information for lexical tones of Mandarin-Chinesee," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2356–2366, 2008.

[28] 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 Pt 1, pp. 2421–2424, 2006.

[29] C. Sanderson and B. C. Lovell, "Multi-region probabilistic histograms for robust and scalable identity inference," in *International Conference on Biometrics*. Springer, 2009, pp. 199–208.

[30] A. Abel, A. Hussain, and B. Luo, "Cognitively inspired speech processing for multimodal hearing technology," in *Computational Intelligence in Healthcare and e-health (CICARE), 2014 IEEE Symposium on*. IEEE, 2014, pp. 56–63.