Using Aerial and Geometric Features in Automatic Lip-reading

Jacek C. Wojdeł, Leon J. M. Rothkrantz

Knowledge Based Systems Group
Delft University of Technology,
Zuidplantsoen 4,
2628BZ Delft, The Netherlands

J.C.Wojdel@cs.tudelft.nl L.J.M.Rothkrantz@cs.tudelft.nl

Abstract

In this paper we present the lip-reading experiments with different sets of the features extracted from the video sequence. In our experiments we use a simple color based filtering techniques to extract the feature vectors from the incoming video signal. Some of those features are directly related to the geometrical properties of the lips (their position and visible thickness). Other features represent the information that relates to the visibility of the other components of the speech production system. The visibility of the teeth and vocal tract for example is described by means of the area they occupy in the image, we call them therefore the aerial features.

1. Introduction

The speech perception by human beings is not restricted to the auditory part of the signal. We perceive the speech as a whole multi-modal stream of information; body movements, gestures and facial expressions, they all influence our speech perception. The modalities other than auditory become crucial in case of hearing disorders or in the extremely noisy environment. However, occurrences such as McGurk effect [1] prove that even normally hearing people in perfect conditions are at least influenced by the visual part of the speech [2].

Those facts suggest that using additional modalities may be beneficial also for automatic speech recognition and/or processing systems. Using a visual part of the speech can be utilized in various ways such as e.g. bimodal speech recognition [3] or visually augmented speech signal enhancement [4].

In lip-reading research it is common use that the geometry of the lips is being extracted from the video sequence and further processed. In that case, the lip-tracking techniques often rely on matching the generic model on the image and extracting the parameters from it [5]. There are however also models that deal not only with the lips' geometry but also the underlying image intensity [6]. We propose here an additional set of features that can complement the geometric features of the lips and allow for better lip-reading results. Both parts of the feature extraction are described in the following sections.

2. Geometric feature extraction

The geometric feature extraction starts from filtering the image using the *lip-selective* filter. The filter must map the given pixel color to the intensity value from $[0, 1]$ interval in such a way that it highlights only the lips in the image.

$$f_{lips} : \langle R, G, B \rangle \in [0, 1]^3 \mapsto I \in [0, 1] \tag{1}$$

Such filters are possible thanks to the fact that lips have usually more reddish color than the rest of the face. The actual form of the filter is not crucial for the way the data is processed further. In our research we use currently a simple *hue-based* filtering (Fig. 1b) or *ANN-based* filtering (Fig. 1c) depending on the quality of the video sequence (see [7]).

As soon as the image filtering is done, the image is transformed into polar coordinates around the center of the mouth. This center point $\langle X_{center}, Y_{center} \rangle$ can be found by computing the center of gravity of the distribution obtained from filtering the image. The resulting intensity function $J(\alpha, r)$ is processed further. There are two interesting properties of this function; its conditional mean $M(\alpha)$ and variance $\sigma^2(\alpha)$ for specific angle.

$$\begin{array}{l} M(\alpha) = \frac{\sum_r rJ(\alpha,r)}{\sum_r J(\alpha,r)} \\ \sigma^2(\alpha) = \frac{\sum_r (r-M(\alpha))^2 J(\alpha,r)}{\sum_r J(\alpha,r)dr} \end{array} \tag{2}$$

Those two values relate directly to the thickness of the lips and their distance from the center of the mouth in a given direction. They describe therefore the shape of the lips in a given video frame. Both of those values can be easily estimated from $J(\alpha, r)$. An example stream of the extracted $\widehat{M}$ and $\widehat{\sigma}$ is shown in the Fig. 2. For a more in depth explanation of this technique see [7].

(a) (b) (c)

Figure 1: Using different lip-selective filters, (a) original image, (b) hue filtered image, (c) ANN filtered image.

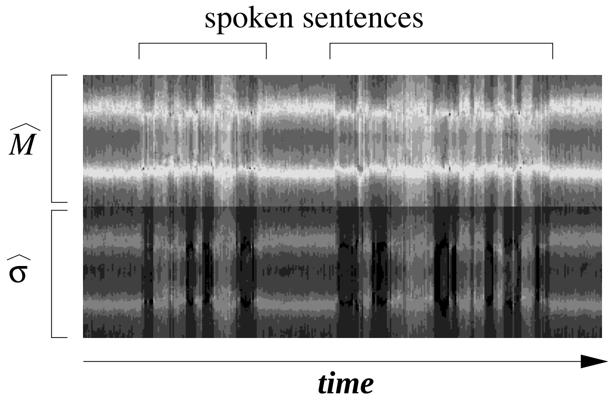

Figure 2: Pairs of $\widehat{M}(\alpha)$ and $\widehat{\sigma}(\alpha)$ vectors extracted from a video-sequence. Those functions after sub-sampling will form the geometric part of the feature vector.

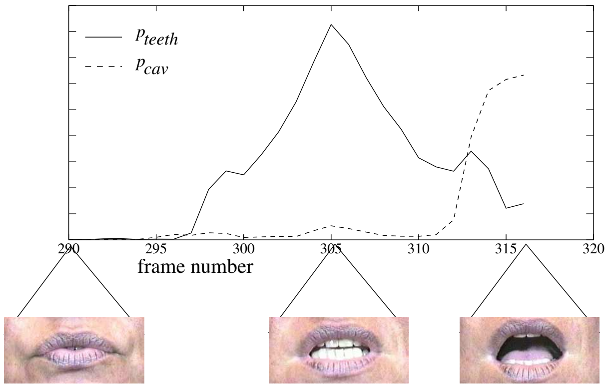

Figure 3: Changes in two of the aerial features in the sequence containing word *zes* (`["zEs]`, six) in frames 290–310 just before the word *acht* (`["Axt]`, eight) beginning in frame 315.

3. Aerial features extraction

The shape of the lips is not the only determinant of the spoken utterance. There are some other important factors such as position of the tongue, teeth etc. Some of them can be observed in the video sequence, the others not. It is essential in case of lip-reading to extract from the visual channel as much information about the utterance being spoken as possible. We propose therefore to use several additional features that complement the geometric information about the shape of the lips.

It would probably be possible to track the actual positions of the teeth and tongue to some limited extent. Such a task would be however too complex and therefore infeasible for a lip-reading application. There are however some easily tractable occurrences that can be measured in the image and which relate to the positions and movements of the crucial parts of the mouth. The teeth for example are much brighter than the rest of the face and can therefore be located using a simple filtering of the image intensity:

$$f_{teeth}(v) = \begin{cases} 0 & if v < t_{teeth} \\ \eta(v - t_{teeth}) & if\ v \geq t_{teeth} \end{cases} \tag{3}$$

The above filter has a steep-wise linear shape and in fact only one parameter; the threshold value t_{teeth}. The slope steepness factor $\eta = (1 - t_{teeth})^{-1}$ is calculated so that the resulting filter produces values in $[0, 1]$ interval.

The visibility and the position of the tongue cannot be assessed as easily as in case of the teeth, especially that the color of tongue is pretty much indistinguishable from the color of the lips. We can however easily assess the amount of the mouth cavity that is not obscured by the tongue. While teeth are distinctly bright, the whole area of the mouth behind the tongue is usually darker that the rest of the face. Therefore we use the following filter to detect it:

$$f_{cav}(v) = \begin{cases} 0 & if\ v > t_{cav} \\ \gamma(t_{cav} - v) & if\ v \leq t_{cav} \end{cases} \tag{4}$$

Both f_{teeth} and f_{cav} filters generate the images with distributions I_{teeth} and I_{cav} respectively. In order to use the information presented in those images, we need to extract some quantitative descriptions of them. We chose to use the total area of the highlighted region and the position of

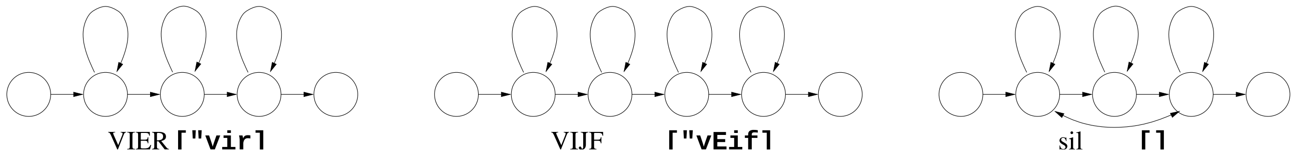

Figure 4: Examples of different length HMMs for different digits.

target						NUL	VIJF	VIJF	ZES	ACHT	ACHT
unconstrained	DRIE	ACHT	VIJF	ACHT	ZEVEN	NUL	VIJF	VIJF	ZES	ACHT	ACHT
constrained					ZEVEN	NUL	VIJF	VIJF	ZES	ACHT	ACHT

Figure 5: The first few digits in the sequence recognized using combined sets of features. The *target* row represents the spoken sequence and the two lower rows show the output from the Viterbi algorithm.

it's center of gravity relative to the center of the mouth:

$$\begin{array}{l} p_\phi = \sum_{x,y} I_\phi(x,y) \\ X_\phi = \frac{\sum_{x,y} xI_\phi(x,y)}{p_\phi} - X_{center} \\ Y_\phi = \frac{\sum_{x,y} yI_\phi(x,y)}{p_\phi} - Y_{center} \\ \phi \in \{teeth, cav\} \end{array} \quad (5)$$

The example changes of the p_{teeth} and p_{cav} are depicted in Fig. 3. In this figure it can be well seen, how the visibility of the teeth dominates during the pronunciation of the word *zes* (["zEs]) and how the increase of the p_{cav} at the end of the sequence relates to the phoneme [A] being spoken later.

4. Experiments

In order to test the feasibility of using the aerial type information in addition to the geometric properties of the lips, we did some experiments with a limited vocabulary recognition. The experiments were run using the Hidden Markov Toolkit (HTK) from Cambridge University on a Sparc Ultra workstation.

4.1. Data acquisition

We made recordings of native Dutch speaker (female) speaking multiple sets containing 10 random digits each. We recorded in total 30 such sets (300 digits). The subject was asked to make pauses between the sets and to speak the digits in a varying pace. There are recordings in which the pauses between words are clearly visible and ones where there are no pauses at all. The recorded video sequences contain only the lower half of the face from nostrils to chin (see Fig. 1a). The movement of the subject's head was constrained only by the fact that she had to read the numbers from the screen in front of her. We recorded the video using a consumer-quality CCD PAL camera and stored digitally in MPEG1 format. In order to reduce the color quality degradation we used a high bit-rate coding.

The obtained data was divided randomly into the training and test sets with test set containing 9 digit sets (90 words). Both sets were then processed in order to extract geometric and aerial features from the video sequences. As the recordings were not done in a single continuous shot; some illumination-induced color variations occurred between different video sequences. This had to be compensated with the recalibration of the used *lip-selective* filter. Fortunately, this happened only a couple of times in the whole recording set. The aerial filters (f_{teeth} and f_{cav}) proved to be much less sensitive to the illumination changes and remained the same for all of the data sets. The resulting data was then prepared for training and recognition using HTK in two separate versions: one containing all features and one stripped off the aerial part of the feature vector.

We used 18 sampling points for geometric features. Together with 6 aerial measurements, the combined data vectors were 42 dimensional. The sampling rate of the signal was usual 25 frames per second as in standard PAL video signal.

4.2. Recognition results

Using HTK we trained 12 Hidden Markov Models (HMMs); one for each digit, one for silence periods at the both ends of the digit sets and one for short pauses between digits. As the digits differ in pronunciation complexity, the models we used differed in number of states. We chose to use a one state per phoneme in each of the models, there was however no enforced correlation between the same phonemes in different digit models. The silence and short pause models had three and one emitting state respectively. Each state was modeled by a single Gaussian distribution. We used in our experiments also the deltas of the consecutive vectors. The training data was not segmented, so the flat start scheme was used for training.

We performed two different versions of testing. In the first case, the number of recognized digits was not constrained by the grammar; the system had no knowledge on how many digits were spoken in the sequence. There is a noticeable amount of insertion errors in this case. It

Table 1: Recognition results obtained for two feature sets, with constrained and unconstrained number of digits.

	%correct	%accuracy	deletions	substitutions	insertions
	unconstrained				
geometric only	80.0	36.7	7	11	39
combined	91.1	60.0	4	4	28
	constrained				
geometric only	75.6	64.4	10	12	10
combined	86.7	81.1	5	7	5

proves that most of those insertions occurs at the beginning and the end of the sequence. That suggests a poorly trained silence model (see Fig. 5).

In the second case we constrained the number of available digits in the sequence. In this way, the system was forced to find the best matching sequence of exactly ten models per digit set. Quite obviously the recognition efficiency in this case improved a bit in comparison to the unconstrained mode.

In Tab. 1 the results of the experiments are summarized. The correct and accuracy percentages given there adhere to the definitions provided in HTK manual:

$$\begin{array}{l} \%correct = \frac{N-D-S}{N} \times 100\% \\ \%accuracy = \frac{N-D-S-I}{N} \times 100\% \end{array} \qquad (6)$$

where N, D, S and I are respectively: total number of words, number of word deletions, number of word substitutions and number of word insertions.

5. Conclusions

We have shown that the addition of the aerial type of features to the geometrical ones can improve lip-reading efficiency. In a simplified example of the limited vocabulary, single speaker lip-reading, the proposed method proves to be reasonably efficient. The overall gain in recognition rate when combining the aerial and geometric features is in the 10-15% range. Moreover, the proposed set of features can be easily extracted from the video sequence without too much computational overhead. The processing model allows for almost real-time operation on available hardware.

There seems to be a problem with training a silence models for lip-reading. Even when not speaking people tend to move their lips, open mouth etc. Training of the silence model proves therefore to be a non-trivial task. It has to be investigated whether using only a single model is sufficient to cover a wide range of possible lip movements that are not speech related. The possible solutions to this problem could be: using several silence models (we would have to differentiate the labeling in this case) or using a multiple Gaussian mixtures in the same model.

Further experiments need to be conducted with the multiple speaker data and in context of continuous lip-reading. The proposed set of processing techniques was designed with the person-independency in mind, but it has to be proven as such yet. We have to consider further also the differing capabilities of visually proper articulation among different subjects.

6. References

[1] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature*, vol. 264, pp. 746–748, 1976.

[2] A. Adjoudani, T. Guiard-Marigny, B. L. Goff, L. Reveret, and C. Benoit, “A multimedia platform for audio-visual speech processing,” in Kokkinakis *et al.* [8].

[3] S. Nakamura, R. Nagai, and K. Shikano, “Improved bimodal speech recognition using tied–mixture HMMs and 5000 word audio–visual synchronous database,” in Kokkinakis *et al.* [8].

[4] L. Girin, G. Feng, and J. Schwartz, “Noisy speech enhancement by fusion of auditory and visual information: a study of vowel transitions,” in Kokkinakis *et al.* [8].

[5] T. Coianiz, L. Torresani, and B. Caprile, “2D deformable models for visual speech analysis,” in Stork and Hennecke [9].

[6] J. Luettin, N. A. Thacker, and S. W. Beet, “Speechreading using shape and intensity information,” in *Proceedings of ICSLP 96*, (Philadelphia, PA, USA), 1996.

[7] J. C. Wojdel and L. J. M. Rothkrantz, “Robust video processing for lipreading applications,” in *Proceedings of Euromedia2001*, (Spain), 2001.

[8] G. Kokkinakis, N. Fakotakis, and E. Dermatas, eds., *Proceedings of ESCA, Eurospeech97*, (Rhodes, Greece), ESCA, 1997.

[9] D. G. Stork and M. E. Hennecke, eds., *Speechreading by Humans and Machines*, vol. 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, (Berlin), Springer Verlag, 1995.