Visual Features for Lithuanian Phone Recognition

G. Tamulevicius[1], D. Eringis[1]
[1]*Recognition Processes Department, Vilnius University Institute of Mathematics and Informatics,*
Akademijos St. 4–307, LT-08663 Vilnius, Lithuania
gintautas.tamulevicius@mii.vu.lt

***Abstract*—Paper presents visual features for speech recognition. Visual speech recognition can be applied as the support for the acoustical recognition process or as a stand-alone speech recognition approach in the case of absence of the acoustical data. Static geometrical features of the speaker‘s lips were used for recognition. Statistical analysis of the feature values enabled us to select individual feature for each analyzed phone and the combined feature set for all phones. The discriminative power was tested using Bayes classification rule. 4th order analysis enabled us to recognize phones with the average rate of 60.3 %, the highest recognition rate of 81.3 % was achieved identifying phone [o].**

***Index Terms*—Visual features, Feature extraction, Speech analysis, Speech recognition**

I. INTRODUCTION

Accurate and reliable speech recognition is the aim for all speech applications and systems. Accurate recognition systems would enable to realize speech-to-text systems, voice controlled services [1], [2]. However, field recognition accuracy is far down [3]–[5] from accuracy in laboratory environment with low noise level, limited number of speakers, correct pronunciation, always the same and incommutable sound equipment, fixed distance to microphone, etc. [3], [4].

The first way to solve these problems is the employment of robust feature system for acoustical signal analysis. Various feature systems and techniques are applied for this purpose – mean subtraction, mel frequency based cepstral analysis, pre-filtering, speaker adaptation. Despite these efforts the acoustical analysis still cannot give robust and reliable speech recognition. The alternative way is the employment of supplementary features, i.e., multimodal speech recognition [2], [6], [7].

The most obvious solution is the use of visual features [6], [8]. Bimodal (audio-visual) recognition of speech is natural for humans thus some linguistic information is carried by speaker’s view. In this case we can talk about visual speech recognition (in opposition to acoustical speech recognition). Visual features characterize various speech relative attributes of speaker’s face: area of region of interest, size / shape of lips, various distances between particular points of lips, appearance of the tongue, teeth, position and velocity of jaw and lips, grey levels of nose, eyes, and mouth windows, etc. [6], [7], [9], [10].

In this paper we shall employ visual features for Lithuanian phone recognition with the aim to determine linguistic content in visual features. Experimental results of recognition of phones are given.

II. AN ISSUE OF LITHUANIAN SPEECH RECOGNITION

Lithuanian is a very ancient and complex language. It is distinguished for its morphological variability, redundant word forms, huge amount of synonyms, etc. These language traits require powerful linguistic processing for reliable continuous speech recognition.

Furthermore, Lithuanian has very similar phones (phonemes) and this fact is essential for correct Lithuanian isolated word recognition. Words “mama” (“mom”) and “mana” (“manna”) are distinguished only by one phoneme for example. And these phonemes (and whole words therefore) are almost indistinguishable acoustically (Fig. 1).

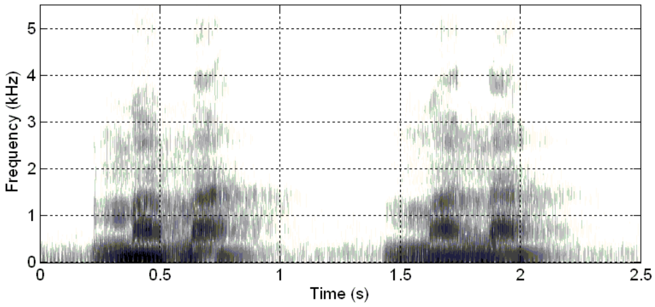

Fig. 1. Spectral diagrams of words “mama” and “mana”.

Very similar situation is with pairs /o/ and /u/, /b/ and /g/, /d/ and /b/, /d/ and /g/. Different words with these phones swapped (we will be talking about phones as higher level of phonemes) are confusing and misleading.

In human interaction the acoustical analysis is supplemented with additional sources of information like visual analysis, context knowledge, and even intuition. And this helps to improve recognition of misleading words. In automatic speech recognition not all means of extra analysis are available and automation of some of them is complicated.

One of the possible solutions is the use of visual features. As we can see in Fig. 2 acoustically indistinguishable phones

Manuscript received May 18, 2012; accepted February 16, 2013.

|m| and |n| can be distinguished visually, i.e. visual configuration of lips is different for phones. In this case view of the lips is the only distinguishable feature and this can be explored for robust recognition.

Hypothetical model of audio-visual speech recognition is the following. Synchronized audio-visual recording of the speech is implemented. Acoustical signal is analyzed and recognized. In the case of unreliable acoustical recognition visual analysis would be started:

1) *Selection of appropriate video frames;*
2) *Detection of lip contours in frames;*
3) *Extraction of visual features;*
4) *Classification of features.*

The results of visual recognition would be compared and aggregated with the results of acoustical recognition giving the final recognition result.

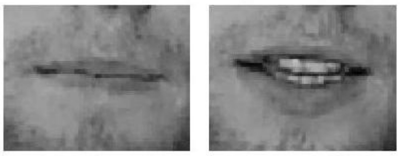
a) b)

Fig. 2. Configuration of lips pronouncing [m] (a) and [n] (b).

In this paper we will explore the use of visual features for classification of misleading Lithuanian phones. We will analyze phonemes /b/, /d/, /g/, /m/, /n/, /o/, and /u/ forming aforementioned pairs. Here we will suppose that acoustically indistinguishable phones will be the case for the visual analysis (acoustical analysis is not performed).

III. Detection of Articulators

To determine contour of lips i.e., the region of interest (ROI), we implemented the edge detection using Canny algorithm [11], [12]. This algorithm detects a wide range of edges by finding the local maxima of the gradient, which is calculated using Gaussian filter's derivative. This approach has a potential of noise attenuation and minimizes multiple detection of the same contour.

Canny algorithm is implemented applying following steps [11]:

1) Detection of real edge points, this is done through maximizing image's signal-to-noise ratio;
2) Localization of edge points;
3) Only a single edge point is selected, the others are repressed;
4) Suppression of non-maximum edges;
5) Hysteresis thresholding.

Firstly, one usually interferes with noises in images, thus one needs to solve signal-to-noise ratio issue. Assume that we have filter with impulse response *f(x)* and edge *G(x)* centered at 0. Filter's response to this edge is calculated by a convolution integral in the range [-*W*, +*W*] (*W* is the boundary values of image pixel *x*)

$$H_G = \int_{-W}^{+W} G(-x) f(x) dx. \tag{1}$$

The root-mean square (RMS) response to noise n(x) is found by following

$$H_n = n_0 \left[\int_{-W}^{+W} f^2(x) dx \right]^{1/2}, \tag{2}$$

where n_0^2 is amplitude of mean-squared noise. Having expressions (1) and (2) we can derive the signal-to-noise ratio equation

$$SNR = \frac{\left| \int_{-W}^{+W} G(-x) f(x) dx \right|}{n_0 \sqrt{\int_{-W}^{+W} f^2(x) dx}}. \tag{3}$$

Localization operator denotes image point which should be as close as possible to the center of current contour. Localization is defined as follows

$$Localization = \frac{\left| \int_{-W}^{+W} G'(-x) f'(x) dx \right|}{n_0 \sqrt{\int_{-W}^{+W} f'^2(x) dx}}. \tag{4}$$

To eliminate multiple edge responses, Canny proposed to find mean distance between zero-crossings of *f'*, it can be written as follows

$$Elimination = \pi \left(\frac{\int_{-\infty}^{+\infty} f'^2(x) dx}{\int_{-\infty}^{+\infty} f''^2(x) dx} \right)^{1/2}. \tag{5}$$

Derivation of (4) and (5) is thoroughly commented in [11].

The non-local maximum suppression step is used to remove non-local maximum points from investigated images, i.e., the algorithm moves along the gradient direction finding pixels with higher gradient value. If such pixel is found it becomes edge candidate, otherwise it is marked as background pixel. Pixels with values higher than threshold *Hw* are included in the edge and pixels with values lower than *Lw* are rejected.

The final step is a lip detection using Haralick-Shapiro method [13], [14]. Neighboring pixels are grouped together into a cluster. As experiments showed the cluster with the biggest area is lip contour usually.

IV. Feature Selection

Every phone has its own vocal tract configuration: position of tongue, shape of lips, openness of nasal cavity, etc. Visually phones distinguish with openness and form of lips, teeth, and tongue position. Vowels are pronounced with open mouth and teeth, with various levels of openness and form of lips. Most of consonants are pronounced with closed lips, some of them have open and close phases. It is obvious that direct relation exists between pronounced phone and external vocal tract configuration.

This fact is exploited in visual speech feature extraction. Various geometrical features of lips contour are extracted and used as visual features of speech. To introduce these features the Fig. 3 with contour of lips is given.

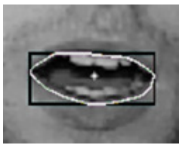

Fig. 3. Extracted lip contour with central point.

The following geometrical features of articulators are extracted and analyzed:

- Lip area (*S*) – number of pixels inside irregular contour (white line in Fig. 3);
- Convex area (*Sc*) – number of pixels inside irregular contour including pixels which are on boundary of the same contour. Lip and convex areas should describe the openness of the lips;
- Length (*L*) of rectangular contour;
- Width (*W*) of rectangular contour;
- Equivalent diameter (*D*) – the diameter of a circle with the same area as irregular contour, it is computed as follows

$$D = \sqrt{4 \cdot S / \pi}; \tag{6}$$

- Extent (*Ex*) – proportion of pixels in irregular contour area and rectangular area

$$Ex = S/(L \cdot W); \tag{7}$$

- Solidity (*Sl*) – the proportion of Lip area and Convex

$$Sl = S/Sc; \tag{8}$$

- Perimeter (*P*) – number of pixels defining irregular contour's length. Perimeter should describe the lip openness too;
- Eccentricity (*Ec*) – scalar defining irregular contour's likeness to circle ($Ec = 0$), ellipse ($0 < Ec < 1$), parabola ($Ec = 1$) or hyperbola ($Ec > 1$). It should be the feature for evaluating form of lips;
- Major axis length (A_{max}) indicates the longest irregular contour's axis which crosses the center of the image;
- Minor axis length (A_{min}) – the shortest irregular contour's axis which crosses the center of the image.

These features are static by nature and will represent the instantaneous geometrical features of the lips. The features will be used without any normalization thus recognition will be speaker and camera position dependent. As these features are extracted from multiple observations of single vocal tract configuration we can expect these features being strongly correlated.

V. Experimental Testing

The speech was recorded using resolution of 640 × 480 pixels, with 30 frames per second rate setting. Face lighting and distance between speaker and camera throughout experiment was maintained in similar level: 70-80 cm. Extracted images were manually cropped to 86 × 68 pixel-size resolution to distinguish approximate position of lips.

Seven phones were selected for experimental testing: /b/, /d/, /g/, /m/, /n/, /o/, and /u/. Visual features were extracted from frames of video stream with the face of the speaker (the records of one male speaker were used in the experiment). Each phone was pronounced 27 times and 96 images for every phone were extracted thus giving 96 visual instances of every phone. The whole feature set was extracted for every phone instance thus making collection of 672 feature vectors.

First of all the statistical analysis of the features was performed. Hypothesis about Gaussian distribution of feature values was tested. We used the Chi-Square goodness of fit test with 95 % confidence level. Analyzed feature values and estimated Chi-Square criteria values did not disprove the hypothesis.

Correlation analysis (with the 95 % confidence level) revealed a high degree of correlation of the features (in some cases a correlation level of 0.99 was obtained) so our presumption about correlation was confirmed.

The discriminative power of features was tested during classification of phones. Two types of classification experiments were performed.

In the first case the task was to recognize the phone in misleading pairs (e. g. to recognize /o/ in pair of /o/ and /u/). The individual feature set for every phoneme pair was formed for this purpose. The features were selected according to their separability – features with high degree of overlap for particular phones were rejected as indistinguishable (for example features *L*, *W*, and *D* were rejected in case of /b/ and /d/ phones). Single feature for one pair was selected for these reasons:

- To eliminate the effect of feature cross-correlation;
- To avoid of "curse of dimensionality". High order feature sets require huge data amounts.

The Bayes classification with Expectation-Maximization evaluation of probability density function was performed.

In order to eliminate the effect of data set size on results, the 3-fold test was executed: feature data was divided into equal 3 subsets and the experiment consisted of 3 tests with every subset as testing (the remaining 2 subsets were used as training data). Table I gives average recognition results.

TABLE I. RECOGNITION RATES OH PHONES.

Phone pair	**Feature**	**1st phone recognition rate, %**	**2nd phone recognitio n rate, %**	**Average recognitio n rate, %**
b-d	*Ec*	94.8	83.3	89.1
b-g	*Ec*	82.3	88.5	85.4
d-g	*Sc*	76.0	88.5	82.3
m-n	A_{min}	89.6	60.4	75.0
o-u	E_x	90.6	78.1	84.4

The highest recognition rate was obtained in case of /b/ phoneme (94.8 %), the lowest was in case of /n/ (60.4 %). The average recognition rate was 83.2 %. The result is high considering the first order analysis.

The practical mean of these results is the following: in the case of vague acoustical recognition we must have information about conflicting phonemes in order to use visual discriminating features effectively. In a real system this information is not always available so we need to form a feature set allowing to recognize the pronounced phone without a prior information.

The second experiment was designed to test the overall phone recognition accuracy. A new feature set was formed –

all the phone-specific features were aggregated into one feature set. This set consisted of eccentricity Ec, convex area Sc, minor axis length A_{min}, and extent Ex.

Again 3-fold experiment was performed in order to get comparable results. Average phone recognition rates with the confidence intervals (evaluated with 95 % confidence level) are given in Fig. 4.

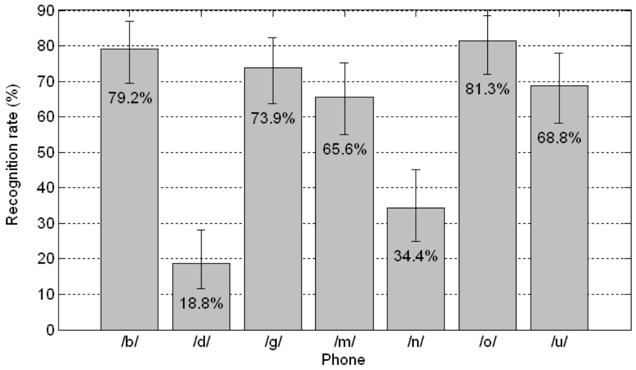

Fig. 4. Average results of phone recognition

The average phone recognition rate 60.3 % was much lower in comparison with recognition in pairs. It is not surprising as we used generalized feature set for all phones. Recognition rate of /d/ and /n/ was lower than average rate and these results coincided with the results of the first experiment where pairs with /d/ and /n/ gave the lowest score. This could be reasoned with high overlap of features Sc and A_{min}.

Similar geometric features (lip contour height, size, area, etc.) were used for various languages phoneme and word recognition [10], [15]–[18]. Phonemes are recognized with the accuracy of 80–90 %. These rates are achieved using high order (up to tens) static and dynamic features. In our case recognition rate was lower but low order of the proposed feature set should be highlighted. 4th feature order allows reducing calculation and storage loads and are much smaller than bitmap or high order geometric features widely used in image processing.

Word recognition rates are much lower and range from 15 % to 40 %. This can be explained by coarticulation effect, many-to-one relationship between phoneme and its configuration of lips, dependence on speaker and camera settings.

To implement robust speaker-independent visual recognition of phones (and speech) some normalizing schemes should be used in order to eliminate speaker distance to camera, image resolution and brightness effect on feature values. Noticeable improvement of recognition rate could be obtained by using dynamic features.

VI. CONCLUSIONS

Following conclusions were made analyzing experimental results:

- The use of geometrical features enables us to reduce the feature order down to 4th thus simplifying analysis and classification task.
- Set of phone-specific features does not ensure the highest overall phone recognition rate. This is conditioned by overlapping and correlation of phone-specific features in aggregated space.
- The phone recognition rate could be improved by adding the new features (e. g. dynamic features). These features should be independent and uncorrelated.

Future research direction is the analysis of the whole phone set and extension of the feature set with the aim to fulfill visual and audio-visual speech recognition.

REFERENCES

[1] R. Maskeliūnas, A. Rudžionis, K. Ratkevičius, "Modeling of call services for public sector", *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering),* no. 4, pp. 81–86, 2010.

[2] J. Kaukėnas, G. Navickas, L. Telksnys, "Human-computer audiovisual interface", *Information Technology and Control*, vol. 35, no. 2, pp. 87–93, 2006.

[3] X. Luo, I. Soon Yann, Ch. Yeo Kiat, "An auditory model for robust speech recognition", *Audio, Language and Image Processing*, pp. 1105–1109, 2008.

[4] G. Čeidaitė, L. Telksnys, "Analysis of factors influencing accuracy of speech recognition", *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering),* no. 9, pp. 69–72, 2010.

[5] D. Balbonas, G. Daunys, "Fonemų klasifikavimas panaudojant garso ir vaizdo informaciją", *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering),* no. 5, pp. 74–77, 2005.

[6] C. C. Chibelushi, F. Deravi, J. S. D. Mason, "A review of speech-based bimodal recognition", *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 23–37, 2002. [Online]. Available: http://dx.doi.org/10.1109/6046.985551

[7] I. Shdaifat, "Design of a visual front end for audio-visual speech recognition", Ph.D. dissertation, Hamburg University of Technology, 2005.

[8] S. Dupont, J. Luettin, "Audio-visual speech modeling for continuous speech recognition", *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000. [Online]. Available: http://dx.doi.org/10.1109/6046.865479

[9] A. Bagai, H. Gandhi, R. Goyal, M. Kohli, T. V. Prasad, "Lip-reading using neural networks", *International Journal of Computer Science and Network*, vol. 9, no. 4, pp. 108–111, 2009.

[10] L. G. Da Silveira, J. Facon, D. L. Borges, "Visual speech recognition: a solution from feature extraction to words classification", in *Proc. of the XVI Brazilian Symposium on Computer Graphics and Image processing*, Oct. 2003, pp. 399–405.

[11] J. Canny, "A computational approach to edge detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.1986.4767851

[12] H. Seyedarabi, W. Lee, A. Aghagolzadeh, "Automatic lip tracking and action units classification using two-step active contours and probabilistic neural networks", in *Proc. of the Canadian Conference on Electrical and Computer Engineering*, May 2006, pp. 2021–2024.

[13] R. M. Haralick, L. G. Shapiro, *Computer and Robot Vision*. Massachusetts: Addison-Wesley, 1992, p. 672.

[14] P. J. Costianes, J. B. Plock, "Gray-level co-occurrence matrices as features in edge enhanced images", in *IEEE 39th Applied Imagery Pattern Recognition Workshop*, Oct. 2010, pp. 1–6.

[15] P. Damien, "Visual speech recognition of modern classic arabic language", in *Proc. of the International Symposium on Humanities, Science & Engineering Research (SHUSER)*, 2011, pp. 50–55.

[16] N. M. Kaynak, Q. Zhi, A. D. Cheok, K. Sengupta, Z. Jian, K. C. Chung, "Analysis of lip geometric features for audio-visual speech recognition", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 34, no. 4, pp. 564–570, 2004. [Online]. Available: http://dx.doi.org/10.1109/TSMCA.2004.826274

[17] P. Singh, V. Laxmi, M. S. Gaur, "n-gram modelling of relevant features for lip-reading", in *Proc. of the International Conference on Advances in Computing, Communications and Informatics*, 2012, pp. 1199–1204.

[18] K. Lu, Y. Wu, Y. Jia, "Visual speech recognition using convolutional VEF snake and canonical correlations", in *Proc. of the IEEE Youth Conference on Information Computing and Telecommunications (YC-ICT),* Nov. 2010, pp. 154–157.