ARCHIVES OF ACOUSTICS
31, 3, 275–288 (2006)

VOWEL RECOGNITION BASED ON ACOUSTIC AND VISUAL FEATURES

P. DALKA, B. KOSTEK*, A. CZYŻEWSKI

Gdańsk University of Technology
Multimedia Systems Department
Narutowicza 11/12, 80-952 Gdańsk
e-mail: bozenka@sound.eti.pg.gda.pl

*International Center of Hearing and Speech
PROKSIM, Kajetany, Warszawa, Poland

(*received January 25, 2006; accepted March 28, 2006*)

The aim of the research work presented is to show a system that may facilitate speech training for hearing impaired people. The system engineered combines both acoustic and visual vowel data acquisition and analysis modules. The acoustic feature extraction involves mel-cepstral analysis. The Active Shape Model method is used for extracting visual speech features from the shape and movement of the lips. Artificial Neural Networks (ANNs) are utilized as the classifier, feature vectors extracted combine both modalities of the human speech. The system is validated with the recordings of speakers that were not used for the lip model creating and for the ANN training. Additional experiments with the degraded acoustic information are carried out in order to test the system robustness against various distortions affecting speech utterances.

Key words: bi-modal automatic speech utterance recognition, phoneme visual and acoustic features, artificial neural networks, Active Shape Model method.

1. Introduction

Bi-modal automatic speech utterance recognition becomes a well-established trend in literature. This is possible due to the well-researched psychophysical background. Categorization of canonical mouth shapes that accompany speech utterances has been carried out and resulted in so-called visual phonemes or "visemes" [1]. Traditionally visemes provide information that complements the phonetic sequence in the case of phoneme confusability [2], however it was discovered lately that video analysis helps much to improve classification accuracy of speech utterances that are hidden by the background noise [3]. Therefore the bi-modal approach to speech unit classification in real conditions is a subject to be still pursued [1–4].

Although significant progress has been made in the domain of automated speech recognition systems, there exist many unsolved problems. Speech recognition systems are usually very sensitive to the background noise and often fail when there are more than one speaker talking simultaneously. Signals that are misclassified can often be corrected with the use of higher level context information, including vocabulary and grammar databases, but it is almost impossible when it comes to the recognition of spelled strings of letters denoting proper names, addresses or other sets of special words. On the other hand, humans have no problem with understanding such words. People subconsciously use additional information from the signal source itself, like positional information about visual articulators (lip movements, teeth and tongue position). In fact, human speech perception is inherently bi-modal [5]. Most visual speech information is contained in the lip contour movements. Also data about the teeth and tongue visibility provide important speech cues [6]. Thus it seems natural to incorporate additional visual information into the speech recognition system. Possible applications of such systems include creating efficient interfaces for working with the computer especially for people with communication sense impairments, amongst them – a convenient way for ticket reservation, booking rooms, automated information centers, etc.

Visemes, or visual speech features are widely used for studying the kinematics of speech production [7] and for speech recognition itself [8]. Mouth modeling is also used for audio visual synchronization [9] and speaker identification [10]. The study presented concentrates on the isolated vowel recognition system combining both acoustic and visual data that may facilitate phoneme training for hearing impaired people. The acoustic part of the system utilizes mel-cepstral analysis [11], which is one of the most common method of speech description and feature extraction. The visual part of the system utilizes model-based approach for extracting speech information from image sequences. Its advantage over the image-based approach is that important features are represented in a low-dimensional space and are normally invariant to translation, rotation, scaling and illumination. Conversely, a disadvantage is that a particular model may not consider all relevant speech information.

The classification process presented in the research conducted by the authors is based on Artificial Neural Networks. ANNs are employed for speech vowel classification based on both visual and acoustic features of the human speech. The important feature of the ANNs is the generalization ability and much less computing complexity of the classification process comparing to other classifiers, for example to the nearest neighbor technique. The goal of the study is to test whether utilizing bi-modal-based feature vectors improves the accuracy of the classification process in case when the background noise is present along with the speech utterance. The assumptions of the system were presented at the 118th Audio Engineering Society Convention and from the results obtained at that stage it could be concluded that combining second modality to the classification process helped the system to increase number of correct answers [12]. The system was also presented at the 7th International Workshop on Mathematical Methods in Scattering Theory and Biomedical Engineering in Nymphaio, Greece, and was published there [3].

For the purpose of experiments two sets of video recordings were prepared. The first one consists of 108 image sequences of nine speakers, each repeating six Polish vowels (a, e, i, o, u, y) twice. Its content reflects a variety of speakers' features, such as different lip shapes, make-up, facial hair and illumination, which makes the classification process more challenging. This set is used to create lip models and to train ANN. The second, validation set consisting of 36 recordings of three speakers is used only to validate results. All recordings are saved in the DV standard (video resolution equal to the full PAL resolution, 25 frames per second, sound sample frequency equal to 48 kHz with 16 bits resolution).

2. Acoustic speech features

In the system Mel Frequency Cepstral Coefficients (MFCCs) [11] were used as the acoustic speech features in the speech recognition system. MFCCs are based on a short-term spectrum. The power spectrum bins are grouped and smoothed according to the perceptually motivated melfrequency scaling. Then the spectrum is segmented into critical bands by means of a filter bank. Finally, a discrete cosine transform is applied to the logarithm of the filter bank outputs resulting in vectors of decorrelated MFCCs features, according to the formula:

$$c_n = \sum_{k=1}^{K} \log X_k \cdot \cos\left(\frac{n\,(k-0.5)\,\pi}{K}\right), \quad n = 1...N, \tag{1}$$

where K is the number of critical bands, N is the desired length of the spectrum and X_k are the filter bank energies passing k-th band-pass filter.

3. Lip models

The system presented in this paper utilizes the Active Shape Models (ASM) [14]. These are flexible models, which represent the boundary or other significant points of an object by a set of labeled points. ASMs use knowledge about possible shape deformation from the statistics of a training set which is labeled by hand. Thereby no heuristic limits for shape deformation are used.

3.1. Model definitions

Two different lip models were studied. The first one (M1) describes outer lip contour while the second one (M2) describes both outer and inner lip contour. These models are invariant against linear conformal image transformations, thus matching the particular lip contour requires additional information regarding translation (t_x, t_y), rotation (Θ) and scale (s). The M1 model consists of 22 points evenly placed on the outer lip contour. The M2 model additionally utilizes 14 points located on the inner lip contour, resulting in 36 points describing the model.

3.2. Training set

The training set used to create the models consisted of 216 images of 9 speakers; two images (the first and the middle frame) were taken from each recording. Thus the training set combined images of nearly closed lips and widely opened ones.

The i-th shape (i = 1...M) consisting of N points can be described as a vector [14]:

$$x_i = \left[\; x_{i1} \;\; y_{i1} \;\; x_{i2} \;\; y_{i2} \;\; ... \;\; x_{iN} \;\; y_{iN} \; \right]^{\mathrm{T}}, \tag{2}$$

where T denotes a transposition.

The lip model contains information about the mean shape and the most common shape deformations, which are acquired from the statistical analysis of the training set. To make this analysis valid, differences in the lip translation, rotation and scale need to be minimized. However simple normalization of each shape separately to the unit width, zero rotation and zero translation could introduce artificial dependencies between shape points that would distort the model. Instead, an iterative algorithm was employed, in which each shape was aligned to the mean shape (in the first step to the freely chosen shape from the training set), and then only the mean shape of all aligned shapes was normalized [14]. Aligning two shapes to each other requires finding the values of translation t_x, t_y, rotation Θ and scale s, which allow to minimize the difference between two vectors describing the shapes. In addition, weights were used to give more significance to those points which are the most stable over the set, i.e. to the ones that move about less with respect to the other points in the shape.

After completion of the aligning process the mean shape and the set of N training shapes are prepared for the statistical analysis.

3.3. Statistical analysis

Statistical analysis of the aligned shapes from the training set employs Principal Component Analysis (PCA) of the covariance matrix, which is given by the equation [15]:

$$S_{2N\times 2N} = \frac{1}{M}\sum_{i=1}^{M} [x_i - \overline{x}][x_i - \overline{x}]^{\mathrm{T}}, \tag{3}$$

where $\overline{x}$ is the mean shape and M denotes number of objects.

The results of PCA of the covariance matrix are $2N \times 2N$ matrix containing the eigenvectors and $2N$ eigenvalues. The eigenvectors are orthogonal and their length is one unit. The eigenvectors with the largest eigenvalues describe the most significant modes of variation. In particular, the variance described by an eigenvector is equal to its corresponding eigenvalue.

A normalized shape can be approximated as follows [16]:

$$x = \overline{x} + \mathbf{P} \cdot b, \tag{4}$$

where $\mathbf{P} = [\, p_1 \;\; p_2 \;\ldots\; p_t \,]$ is the matrix of the first t $(t < 2N)$ eigenvectors corresponding to the largest eigenvalues and b is a vector containing weights for each eigenvector. The number of eigenvectors used to describe the main modes of shape variation is usually much smaller than the number of elements in the shape describing vector.

4. Locating lips

In order to utilize shape models to locate lips in the image sequence, it is necessary to measure the fit between the shape model and lip contour in the image. Thus two luminance models (one for each shape model), describing grey-level appearance around the lip contour, were created.

For every image in the training set (identical to the one used for shape model acquisition) N luminance profiles were calculated. Profile g_{ij} is a vector containing N_p greylevel values evenly distributed on the sector perpendicular to the lip contour and anchored to the j-th point of the i-th image (Fig. 1). These profiles were appended to each other to create one global luminance profile h_i for the i-th image in the training set. Thus each global profile is a vector containing $N \cdot N_p$ elements [16]:

$$h_i = \left[\; g_{i1} \;\; g_{i2} \;\; \ldots \;\; g_{iN} \;\right]^{\mathrm{T}}, \qquad i = 1...M. \tag{5}$$

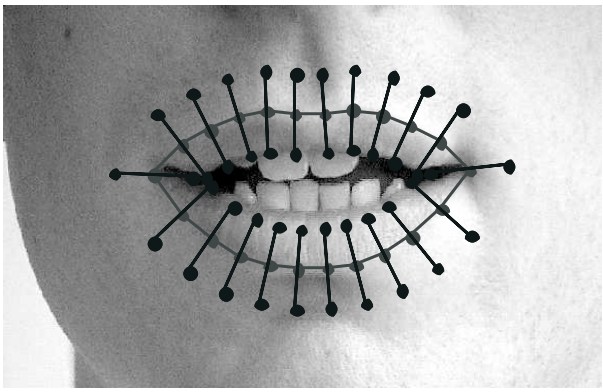

Fig. 1. Luminance profile acquisition for the M1 model.

In analogy to the lip modeling, the mean global luminance profile $\overline{h}$ and the covariance matrix are calculated, and the *PCA* analysis is performed. Any luminance profile can be in this way approximated according to the formula:

$$h = \overline{h} + \mathbf{P}_g \cdot b_g, \tag{6}$$

where $\mathbf{P}_g = [\, p_{g1} \;\; p_{g2} \;\ldots\; p_{gt} \,]$ is the matrix of the first t $(t < N \cdot N_p)$ eigenvectors corresponding to the largest eigenvalues and b_g denotes a vector containing weights for each eigenvector.

Locating lips in the image is an iterative process. Its goal is to minimize the cost function describing the fit between the luminance model and the image. Having a lip shape x and its position in the image, weights b_g of the eigenvectors of the luminance profile h corresponding to the shape x can be found using the formula:

$$b_g = \mathbf{P}_g^{\mathrm{T}} \left(h - \overline{h}\right). \tag{7}$$

The cost function E_p is defined as the mean squared error between the luminance profile h and the luminance model:

$$E_p = \left(h - \overline{h}\right)^{\mathrm{T}} \left(h - \overline{h}\right) - b_g^{\mathrm{T}} b_g. \tag{8}$$

The smaller cost function value E_p, the greater similarity between the luminance profile h and the statistical luminance model derived from the analysis of the training set (the $\overline{h}$ vector and the $\mathbf{P}_g$ matrix). Thus the smaller E_p value, the more accurate aligning between the shape x and lip contour in the analyzed image.

The Downhill Simplex Method [13] was employed to minimize the value E_p of the cost function. The algorithm uses translation parameters t_x and t_y, scale s, rotation Θ and the vector of shape parameters b as variables for the multidimensional optimization process. This algorithm is commonly used in such applications and is effective in a sense of result achievement. In addition, the algorithm result has a very convenient geometric interpretation, hence it is very suitable for analysis. Conversely, it is time-inefficient because it requires value of the optimized function to be calculated more times than other algorithms do.

It is assumed that the region of the image containing lips is known. The optimization algorithm is initialized with the mean shape ($b = 0$) placed randomly in this region. In the next step, the b_g vector of the luminance profile is calculated (7) and the value E_p of the cost function is computed (8). The algorithm iteratively changes the lip shape (b vector) and its location in the image (t_x, t_y, s and Θ variables) to minimize the E_p value. The algorithm stops when the difference between subsequent E_p values drops below the assumed threshold.

During image search, each shape mode is restricted to stay within ± 3 standard deviations, which accounts for 99% of variation.

When lips are being located in the image sequence, the lips position in the previous frame is used as the initial estimate in the current one.

5. Experiments

Based on initial experiments, 17 shape modes were used for the M1 model, and 28 shape modes were used for the M2 model, which covers 99.9% of shape variation in both cases. The dimension of the luminance profile N_p was 21. For the M1 model nine luminance modes were used and for the M2 model – 14 modes, which accounts for 80% luminance variation.

The results of the lip locating are largely dependent on the number of luminance modes used. If it is too small or too great, the algorithm of E_p value minimization stops prematurely in the local extremum, resulting in the erroneous lip location.

5.1. Locating lips

There is one major drawback regarding the lip locating experiments. This regards the lack of an objective, automatic method of the algorithm result evaluation. This task requires all output sequences with the lip position and shape marked to be viewed one by one, thus the evaluation is subjective and, although the strict criteria were introduced, the same result of the lip locating could be judged differently. Table 1 contains results of lip locating with both shape models.

Table 1. Lip locating results.

Rating	Model M1		Model M2	
	No. of recordings	[%]	No. of recordings	[%]
3 (best)	25	23.2	22	20.4
2	57	52.8	36	33.3
1	23	21.3	38	35.2
0 (worst)	3	2.8	12	11.1
3+2	82	75.9	58	53.7
total	108	100	108	100

A recorded sequence was evaluated as the whole rather than its frames separately. Recordings in which lip shape was perfectly located in every frame were rated with the number 3. When the extracted lip shape differed slightly from the real shape in one or more frames, but the differences were meaningless in terms of speech recognition, the entire sequence was rated with the number 2. However, when these differences were larger and could have a bad influence on the speech recognition, the recording was rated with the number 1. The lowest ranking, 0, was given the recordings containing at least one frame with a serious lip locating error. All image sequences rated with 2 and 3 were considered correct and only these recordings were used in the speech recognition experiments. Figure 2 presents a few results of the lip locating.

The results of lip locating, especially for the model M1, are satisfactory. Mistakes in the lip shape or location very frequently appear only in the few adjacent frames; according to the assumed criteria such recordings were ranked low. If each frame was evaluated separately, the results would be much better, but this assumption would be incorrect in the terms of speech recognition where lip shapes in the adjacent frames form an indivisible whole.

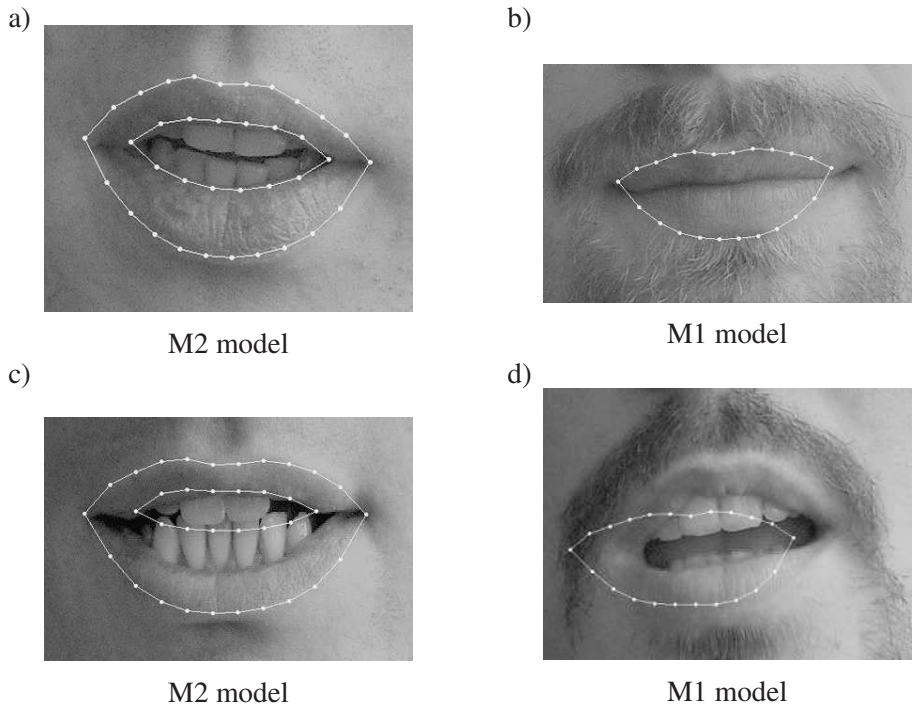

Fig. 2. Results of the lip locating algorithm; typical frames from the recordings rated a) 3, b) 2, c) 1, d) 0.

Lip locating results differ significantly for each vowel. The lowest percentage of ratings 2 and 3 was assigned to the recordings of vowels i and u localized with the M1 model (approx. 70% of all recording of these vowels) and to the vowels a and e localized with the M2 model (nearly 40% of recordings). The highest quantity of ratings 2 and 3 was assigned to the vowel o (94% of recordings for the M1 model and 67% for the M2 model).

Table 2. Lips locating results for recordings from the validation set.

Rating	Model M1		Model M2	
	No. of recordings	[%]	No. of recordings	[%]
3 (best)	7	19.4	6	16.7
2	21	58.3	13	36.1
1	7	19.4	14	38.9
0 (worst)	1	2.8	3	8.3
3+2	28	77.8	19	52.8
total	36	100	36	100

Additional experiments were carried out to test the lip locating algorithm with the second set of video sequences, which consists of recordings that were not used to create lip models. Results obtained are presented in Table 2.

Comparing data in Tables 1 and 2 it may be seen that results of lips locating in both sets of recording are very similar. However there is a slight loss of the best rated results in favour of the less rated ones. This proves that the lip locating algorithm is generally robust against diversity of lip shapes originating from the anatomical features and the way of speaking.

5.2. Vowel recognition

5.2.1. Visual features

Based on the results of lip locating, only image sequences rated with 2 and 3 were used in the phoneme recognition experiments. A three-layer feed-forward artificial neural network (ANN) was employed for this purpose. A matrix, containing feature vectors calculated in the 20 points of the time evenly spaced during the utterance, formed an input to the ANN. To make the results of speech classification robust against the changes in the utterance duration, an interpolation was used to derive the feature vectors. Furthermore, from every recording five sequences shifted by the multiplicity of 1/25 s were derived in order to make the ANN invariant to the time shifting.

The initial stage of experiments started with the training phase of the ANN. Vectors (matrices) of parameters were randomly divided into two sets: training and testing vectors. Each set contained 50% of recordings of every vowel (if a number of recordings taken into account for a vowel was odd, the training set was complemented by more vectors). The error back-propagation algorithm was used to train the ANN. The process of training was considered as finished either when the value of the cumulative error of network responses for the training set of vectors had dropped below the assumed threshold value or when the cumulative error of network responses for the validation set of vectors had been rising for more then 10 cycles in a row. The amount of neurons in the output layer was equal to the number of recognized vowels. The vowel, classified by the ANN, was determined by the highest value of the output signals of neurons in the output layer.

The lip locating algorithm delivers two sets of parameters that can be used as the visual speech features. Lip shape parameters b are the first set, and luminance profile parameters b_g are the second set. The results of vowel recognition are presented in Table 3. The number of recordings given in Tables 3, 4 and 5 for each model constitutes approximately half of recordings which were rated with 2 or 3 points during the localization process, multiplied by the number of sequences derived from every recording.

The M2 model, which more accurately expresses the lip shape, achieved better results than the M1 model. Results of speech recognition with the b_g parameters describing the grey-level appearance around the lip contour turned out to be significantly better than the b parameters describing only the lip shape, because the former parameters

Table 3. Results of vowel classification based on the visual features.

Model	No. of recordings	Features		
		b	b_g	$b + b_g$
M1	200	31.50%	51.00%	45.00%
M2	135	56.30%	69.63%	77.78%

Table 4. Detailed results of vowel classification with the M2 model and $b + b_g$ parameters.

Vowel	No. of recordings	No. of errors	Effectiveness [%]
a	15	5	66.67
e	15	10	33.33
i	25	8	68.00
o	30	2	93.33
u	30	0	100.00
y	20	5	75.00
total	135	30	77.78

contain additional information about the tongue position and the teeth visibility. Furthermore, using both sets of parameters with the M1 model causes deterioration of the results. This enables to conclude that inner lip contour is much more important than the outer one and adding this complementary information to the b_g parameters increased the effectiveness of classification with the M2 model to nearly 80%. Detailed results of six vowels classification with the M2 model and $b + b_g$ parameters are presented in Table 4.

Results of the vowel recognition are as expected. The majority of mistakes were made during the classification of vowels, which are practically impossible to separate based on the visual speech features only. Amongst them, there are such pairs, as a–e and i–y which were very often confused with each other. Additional experiments with the recognition of vowel pairs were also carried out. For a–e pair the ANN correctly classified 76.67% of recordings, and for i–y pair – 95.56%. Every other pair of vowels was separated without mistakes.

5.2.2. Combining both modalities of speech

Ten mel-frequency cepstral coefficients (MFCC) calculated every 10 ms were used as the acoustic speech features. Results of classification with the acoustic only and both acoustic and visual features are presented in Table 5.

The results of vowel classification with the acoustic features are much better that those with the visual features. However combining both modalities of human speech further improves the effectiveness by 1.5% for the model M1 and almost 5% for the model M2.

Table 5. Results of vowel classification based on visual and acoustic features.

Model	No. of recordings	Acoustic features	Visual and acoustic features
M1	200	94.63%	96.00%
M2	135		99.26%

Table 6 contains results of vowel recognition for the recordings of speakers from the second, testing set, which were not used during the ANN training phase. The number of validation recordings given in Table 6 is the number of recordings which were rated with 2 or 3 points during the localization process, multiplied by the number of sequences derived from every recording. Comparing data in this table with Tables 3 and 5 one can notice that the effectiveness of classification based on visual features is very similar. However results of classification with acoustic features are lower by approximately 4%, which means an equivalent loss of the effectiveness of recognition based on both types of features. This proves that visual features are truly speaker-independent while the results of speech recognition based on acoustic features (and both acoustic and visual ones) seem to be related to the particular speaker.

Table 6. Results of vowel classification based on visual and acoustic features for recordings from the validation set.

Model	No. of recordings	Visual features $(b + b_g)$	Acoustic features	Visual and acoustic features
M1	140	47.14%	90.21%	92.14%
M2	95	75.79%		95.79%

Additional experiments were carried out to examine the robustness of systems based on both visual and acoustic features and on acoustic features only against an audio distortion. An extract of a single person continuous speech mixed into the recordings was used as the distortion. Figure 3 shows the results of classification against different distortion level values measured as the total RMS difference in dB between the distortion and the signal.

Effectiveness of the system based on the acoustic features begins to drop when distortion exceeds the level of -20 dB (ten times lower than the level of speech). However the system based on both acoustic and visual speech features remains robust against the distortion until its level reaches the level of speech. When the distortion level is very high (over 100 times higher than the level of speech) the system using the M2 model and both modalities of speech correctly classifies much more recordings (62%) than the system using the M1 model (44%). It proves that the M2 model carries more information about the lip appearance and therefore the system based on it is more robust against audio signal distortions.

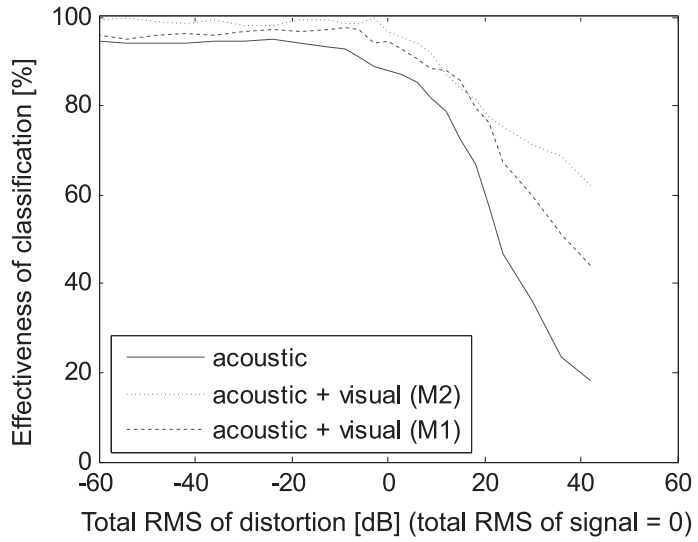

Fig. 3. Results of classification based on acoustic and visual features against different levels of sound distortion.

6. Conclusions

As shown in the study presented automatic vowel recognition process can benefit from adding visual features to FVs. This is especially important in case of real conditions when speech utterance is presented against the background noise. This means that visual information helps to attain greater robustness of the system against distortions in the audio signal, and this can improve the reliability and capacity of the classification process.

A system utilizing only visual information correctly classified nearly 80% of speech samples. This result is very satisfying taking into account a huge similarity between lip movements during articulation of some vowels and a great diversity of lip shapes originating from the anatomical features and the way of speaking. However further improvements of the lip locating algorithm are required to make its results more accurate. This could be achieved by increasing the number of images included in the training set and by increasing the number of speakers in the database. Also it is necessary to develop new parameters based on the known lip shape and location that would more precisely describe the visual modality of human speech.

Results of classification based on the acoustic information are much better than the ones based on the visual information. However, utilizing both modalities in the speech recognition system improves further its effectiveness. It should be remembered that while measuring the performance of a classification system the most difficult are the last percentage points to achieve, thus obtaining accuracy over 90% may be judged as satisfying.

In literature numerous potential applications of systems for extracting and analyzing visual speech features are seen, this especially concerns the field of audiology and logopedics. They can be used to detect various voice and speech defects, e.g. dysarthric speech. Furthermore, they may facilitate speech training for people with hearing and speech impairments, for example as an application enabling to perform interactive phoneme articulation lessons designed particularly for children. In addition, the presented solution, when fully developed, can be used in the speech recognition system designated for hearing impaired people, such an application can especially be useful in the noisy environment where the access to the acoustic information is limited or its quality is unacceptable.

Acknowledgments

This work was supported by the Polish Ministry of Education and Science within the research project No. 3T11E02829.

References

[1] SAENKO K., LIVESCU K., GLASS J., DARRELL T., *Production domain modeling of pronunciation for visual speech recognition*, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pp. 473–476, Philadelphia, March 2005.

[2] VERMA A., FARUQUIE T., NETI C., BASU S., SENIOR A., *Late integration in audio-visual continuous speech recognition*, Proceedings of Automatic Speech Recognition and Understanding, Colorado, 12–15 December 1999.

[3] KOSTEK B., DALKA P., CZYŻEWSKI A., *Audiovisual speech recognition for training hearing impaired patients*, Nymphaio, Sept. 2005, World Scientific Publishing, 2006.

[4] MOTLICEK P., CERNOCKY J., *Multimodal phoneme recognition of meeting data*, Vol. 3206/2004, Text, Speech and Dialogue: 7th International Conference, TSD 2004, Brno, Czech Republic, September 8-11, 2004. Proceedings, Lecture Notes in Computer Science (Sojka P., Kopecek I., Pala K. [Eds.]), 2004.

[5] DODD B., CAMPBELL R., *Hearing by eye: The psychology of lipreading*, Lawrence Erlbaum Press, 1987.

[6] SUMMERFIELD Q., *Lipreading and audio-visual speech perception*, Phil. Trans. R. Soc. Lond., B 335, 71–78 (1992).

[7] BORGHESE N. A., FERRIGNO G., REDOLFI M., PEDOTTI A., *Automatic integrated analysis of jaw and lip movement in speech production*, J. Acoustical Soc. of America, **101** 1, 482–487 (1997).

[8] KUBANEK M., *Audio-visual recognition of Polish speech based on hidden Markov models* [in Polish], PhD Thesis, Czestochowa University of Technology, 2005.

[9] FABIAN P., BADURA S., LESZCZYŃSKI M., SKARBEK W., *Mouth modeling by local PCA for audio visual synchronization* [in Polish], XI International AES Symposium – The Art of Sound Engineering ISSET, pp. 79–85, Kraków 2005.

[10] KUBANEK M., *Method of speech recognition and speaker identification with use audio-visual of Polish speech and hidden Markov models*, Proc. Advanced Computer Systems – Computer Information Systems and Industrial Management Applications, ACS-CISIM, Ełk 2005.

[11] DAVIS S. B., MERMELSTEIN P., *Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences*, IEEE Trans. on Acoustics, Speech & Signal Processing, **28,** 4, 357–366 (1980).

[12] KOSTEK B., DALKA P., *Combining visual and acoustic modalities to ease speech recognition by hearing impaired people*, 118th Audio Engineering Society Convention, Paper No. 6462, Barcelona 2005.

[13] NELDER J., MEAD R., *A simplex method for function optimization*, Computing Journal, **7,** 4, 308–313 (1965).

[14] COOTES T., TAYLOR C., COOPER D., GRAHAM J., *Active shape models – their training and application*, Computer Vision and Image Understanding, **61,** 1, 38–59 (1995).

[15] JACKSON J., *A user's guide to principal components*, John Wiley and Sons, Inc., 1991.

[16] LUETTIN J., THACKER N., *Speechreading using probabilistic models*, Computer Vision and Image Understanding, **65,** 2, 163–178 (1997).