Visual Speech Analysis, Application to Arabic Phonemes

Fatma Zohra Chelali,
Speech communication and signal processing laboratory
Electronics and computer Science Faculty
Houari Boumedienne University of sciences and Technologies,USTHB
Box n°:32 El Alia, 16111, Algiers, Algeria

Khadidja Sadeddine
Speech communication and signal processing laboratory
Electronics and computer Science Faculty
Houari Boumedienne University of sciences and Technologies,USTHB
Box n°:32 El Alia, 16111, Algiers, Algeria

Amar Djeradi
Speech communication and signal processing laboratory
Electronics and computer Science Faculty
Houari Boumedienne University of sciences and Technologies,USTHB
Box n°:32 El Alia, 16111, Algiers, Algeria

ABSTRACT
The aim of this work is to introduce a primary research on Arabic audiovisual analysis. Each language has multiple phonemes and visemes and each viseme can have multiple phonemes. The first part focuses on how to classify Arabic visemes from still images, whereas the second part shows the variation of Pitch for each viseme. We haven't taken co-articulation of visemes in context.

To evaluate the performance of the proposed method, we collected a large number of speech visual signal of ten Algerian speakers male and female at different moments pronouncing 28 Arabic syllabuses. In our work, we demonstrate 11 final visemes representing the 28 consonantal Arabic phonemes.

Keywords
Arabic visemes, speech recognition, audiovisual analysis, pitch.

1. INTRODUCTION
Automatic recognition of audio-visual speech introduces new and challenging tasks compared to traditional, audio-only ASR: In addition to the usual audio front end (feature extraction stage), visual features that are informative about speech must be extracted from video of the speaker's face. This requires robust face detection, as well as location estimation and tracking of the speaker's mouth or lips, followed by extraction of suitable visual features [5, 16].

Visual speech cues play an important role in Human speech perception, especially in noisy environments. Phenomena such as the perceptive illusions of McGurk (speakers confronted to an auditory stimuli /ba/ and a visual stimuli /ga/ perceive the stimuli /da/) or the "Cocktail party" effect (attention centered on a special speaker surrounded by multiple speakers discoursing at the same time) show the significance of visual information in speech perception [19].

In this context, many works in the literature have proved that movements of the mouth can be used as one of the speech recognition channels. Recognizing the content of speech based on observing the speaker's lip movements is called 'lip-reading'. It requires converting the mouth movements to a reliable mathematical index for possible visual recognition [15].

Visemes and phonemes can be used as the basic units of visible articulatory mouth shapes [1]. A phoneme is an abstract representation of a sound, and the set of phonemes in a language is defined as the minimum number of symbols required to represent every word in that language [2].

The set of visemes in a language is often defined as the number of visibly different phonemes in that language. A simple definition of viseme would be that a viseme could be generated from a set of archetypal sounds in a language based on the phonemes of that language, Möttönen, Olivés et.al defines a viseme set as phoneme realizations that are visually inextricable from each other. In lip reading studies viseme categories can be defined as clustering response distributions to observed phoneme articulations [3].

Several standards for visemes [4] exist but they were developed for English language. Since each language comprises of a different phonetic set (sounds) therefore, visemes have to be identified separately for each language. Möttönen, Olivés et.al. developed a Finnish talking head, which had to identify Finnish visemes for each Finnish phoneme [5, 6].

To date, there has been no precise definition for the term, but in general it has come to refer to a speech segment that is visually contrastive from another [3, 6].

It is also important to point out that that the map from phonemes to visemes is also one-to-many: the same phoneme can have many different visual forms. This phenomenon is termed coarticulation, and it occurs because the neighboring phonemic context in which a sound is uttered influences the lip shape for that sound [6, 16].

The purpose of our work is to introduce an Arabic viseme system based on visual speech database through statistical and neural analysis. Experiments were carried in speech communication and signal processing laboratory, faculty of electronics and computing of Algiers, ALGERIA.

The paper is organised as follows. In Section 2, a brief description of the Arabic language is presented. Section 3 details the Arabic phoneme database acquisition used in the experiments, this section presents the theory of visemes developed in our work. Section 4 discusses the results obtained. Finally, Section 5 closes with conclusion and also suggests future work in this area of research.

2. ARABIC LANGUAGE

Arabic is a Semitic language, and it is one of the oldest languages in the world. It is the 5th widely used language nowadays. Standard Arabic has 35 basic the Pharyngealized (L) which is rarely used) and six are vowels, three long and three short [9].

Vowels are relatively easy to identify because of their high energy resulting in instance formant patterns. The vowels are voiced sounds. Arabic phonemes contain two distinctive classes, which are named pharyngeal and emphatic phonemes. These two classes can be found only in Semitic languages. The allowed syllables in Arabic language are: CV, CVV, CVC, CVVC, CVCC and CVVCC where V indicates a (long or short) vowel while C indicates a consonant. Arabic utterances can only start with a consonant [9].

Arabic syllables can be classified as short or long. The CV type is a short one while all others are long.

Table 1. Arabic syllable patterns

	Open	Closed
Short	CV	
Long	CVV	CVC, CVVC, CVCC, CVVCC

3. DESCRIPTION OF THE SYSTEM

3.1 Phoneme database

The corpus is a repetition of the 28 Arabic phonemes spoken by ten native Algerian speaking male and female subjects. A database of facial and speaker features is assigned for one of each known individuals. The recordings were made on camera canon video with 25 frames of size 576*720pixels per second. Then, the data were transferred into computer through IEEE1394 card. The corpus consists of 20 repetitions of every syllabus (phoneme with short vowel) produced by each speaker, 20 still images in format bmp and 20 repetitions for audio file in format wav.The database includes 2800 face images from ten (10) different subjects. The speech signals are acquired during different sessions with a sampling frequency of 22 KHz.

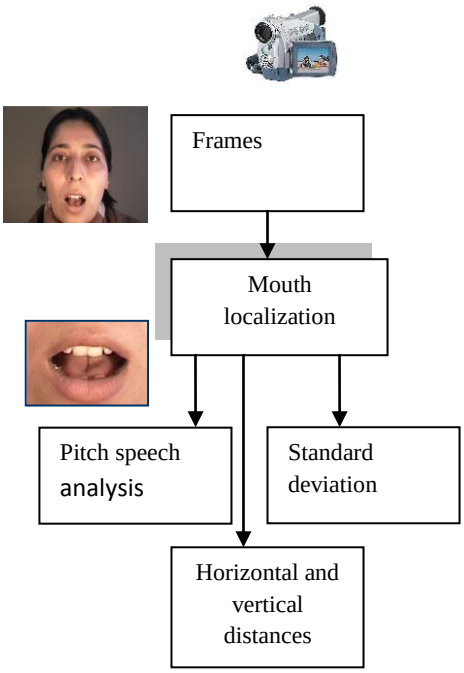

Figure 1. Overview of system acquisition

3.2 The consonant visemes

In lip-reading, audiovisual recognition, speech/speaker recognition, visemes are used as basic speech units. A viseme is a group of phonemes, whose visual expression is the same. For example phonemes p, b, and m create one viseme.

To date, there has been no precise definition for the term, but in general it has come to refer to a speech segment that is visually contrastive from another. For example, aba /?b?/ and ama /?m?/ are two bilabial signals which differ only in the fact that the former is voiced while the latter is voiceless. This difference, however, does not manifest itself visually, and hence the two phonemes should be placed in the same visemic category.

In our work, we investigate visemes representing the 28 consonantal Arabic phonemes using two methods, the statistical parameters (standard deviation) for lip image, the geometrical parameters for the internal and the external lip contour. A MLP neural network is then used for classification.

The following figure shows the lip image for phoneme “A” pronounced by two male and female speakers.

Figure 2. Visual representation of phoneme A

For each frame the lip area is manually located with a rectangle of size proportional to 120*160 and centered on the mouth, and converted to gray scale (figure 2). Finally, the mean and the standard deviation of the values of the pixels of the lip area (of size n*m) are computed by using 20 images for each phoneme sequence to classify our visemes.

$$\bar{I} = \frac{1}{n*m}\sum_{i,j} I(i,j) \quad (1)$$

$$\sigma_I = \sqrt{\frac{1}{n*m}\sum_{i,j}(I(i,j)-\bar{I})^2} \quad (2)$$

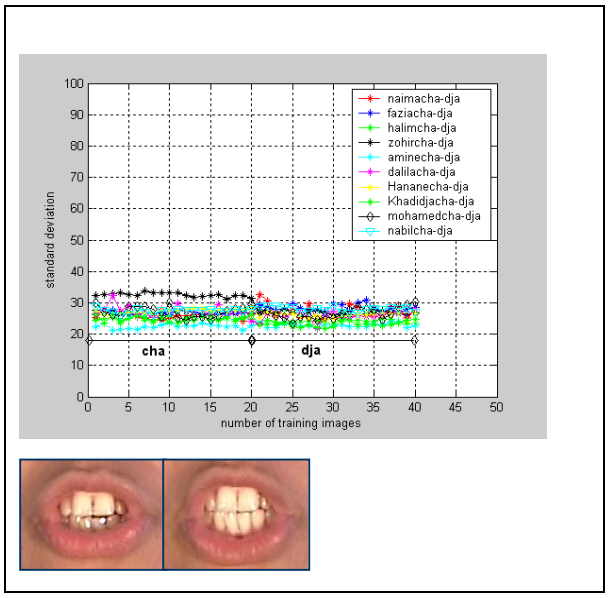

Figure3. The standard deviation for 2nd viseme /dja, cha/ {ش, ج}

We use a set of geometrical parameters based on the internal and external lip shape of the speaker. Several studies have underlined a certain correlation between these geometrical parameters, which allows us to reduce their number to only two corresponding to the height and the width of the internal lip contour. In spite of this, we have chosen to add the parameters of the external contour as mentioned in the work of benoit &al and L.Revéret [12,13,14].

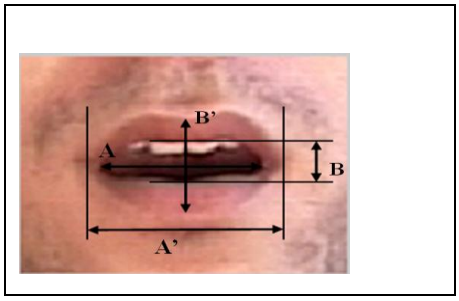

Figure3. The geometrical parameters for visemes classification

A: the width of the internal lip contour (the first horizontal distance).

A': the width of the external lip contour (the second horizontal distance).

B: the height of the internal lip contour(the first vertical distance).

B': the height of the external lip contour(the second vertical distance).

Each image is presented as a matrix containing the four (4) values of the horizontal and vertical distances.

3.3 Pitch analysis

The pitch determination is very important for many speech processing algorithms, a commonly used method to estimate pitch (Fundamental frequency) is based on detecting the highest value of the autocorrelation function in the region of interest. Our perception of pitch is strongly related to periodicity in the waveform in the time domain [20].

In practice, we need to obtain an estimate the autocorrelation $\hat{R}[m]$ from knowledge of only N samples. The empirical autocorrelation function is given by :

$$\hat{R}[m] = \frac{1}{N}\sum_{n=0}^{N-1-|m|}(w[n]x[n]w[n+|m|]x[n+|m|]) \quad (3)$$

where $w[n]$ is a window function of length N.

So that we can find the pitch period by computing the highest value of the autocorrelation.Which has a maxima for m=lT0 [20].

Since pitch periods can be as low as 40Hz (for a very low-pitched male voice) or as high as 600 Hz (for a very high-pitched female or child's voice), the search for the maximum is conducted within a region [20].

4. Experimental results

4.1 The 11consonant Arabic visemes

In our work, we demonstrate 11 final visemes representing the 28 consonantal Arabic phonemes using two methods, the statistical parameters (mean and standard deviation) for lip, the geometrical parameters for the internal and the external lip contour. The pitch analysis is studied in order to demonstrate the variation of temporal and spectral characteristics for the same viseme.

We analyse the standard deviation for the twenty (20) images for each phoneme, figure3 shows the standard deviation for the the first and the second (dja-cha)viseme

We analyse also the geometrical parametrs described in section3. Each image is presented as a matrix containing the four (4) values of the horizontal (HD) and vertical (VD) distances stored in a file (.mat). These parameters are used to classify our visemes as indicated in the following figure.

The following figure shows the two parameters calculated for ten (10 images) for each phoneme (ba and ma), we are interested only on the two distances A' and B' for the first viseme,the distance B for the internal contour is not significant.

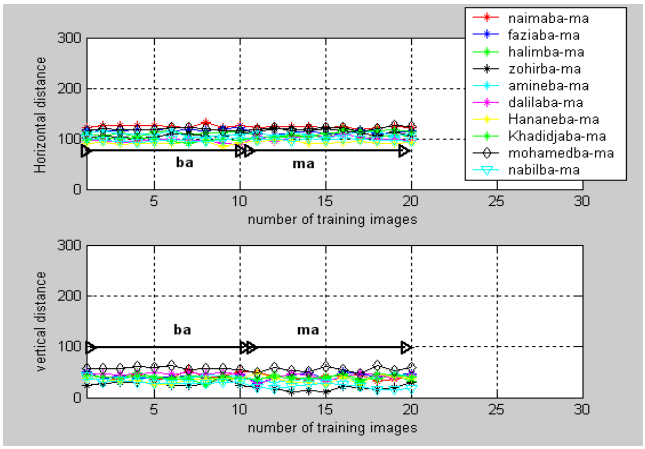

Figure4. Variation of horizontal and vertical distance of the first viseme

In Arabic language, approximately 11 visemes(9visemes described by table 3 and the two short vowel /i/ and /u/) are distinguished.

Table2. Arabic visemes

Viseme	phoneme	API
1. Ba	{ب, م }	/b/, /m/
2. Cha	{ش, ج}	/ š / , / z/
3. dha	{ ث, ذ, ظ)	/ θ/,/ δ/, / δ/
4. Ha	{ أ , ع , غ , ح , ه, ك , خ, ق }	/ʔ, ʕ, γ, ħ, h, k, x, q/
5. Ta	{د , ن , ت , س , ز , ط , ص , ي , ض }	/d,n,t,s,Z, t', ś,j, d'/
6. Fa	{ف}	/F/
7. La	{ل}	/l/
8. Ra	{ر}	/r/
9; wa	{و}	/ w/
10	ظمه ممالة	/u/
11	كسرة ممالة	/i/

In order to better understand the structure of visual speech data and hence to determine the visemes for further automatic lipreading, we use hierarchical classification using the k nearest neighbour.

The vectors are viewed as points in the l-dimensional space and the clusters are described as "continuous regions of this space containing a relatively high density of points, separated from other high density regions by regions of relatively low density of points. We try to give some definitions for clustering [17].Let X be our data set, that is,

$$X = \{x_1, x_2, ..., x_N\} \quad (4)$$

We define as an m-clustering of X, R, the partition of X into m sets (clusters), $c_1, ..., c_m$, so that the three following conditions are met:

$$c_i \neq \phi, \; j = 1.....m \quad (5)$$

$$\cup_{i=1}^{m} c_i = X$$
$$c_i \cap c_j = \phi \quad , i \neq j, i, j = 1.....m \quad (6)$$

In addition, the vectors contained in a cluster ci are "more similar" to each other and "less similar" to the features vectors of the other clusters [17].

However, not all clustering algorithms are based on proximity measures between vectors. For example, in the hierarchical clustering algorithms one has to compute distances between pairs of sets of vectors of X. In the sequel, we extend the preceding definitions in order to measure "proximity" between subsets of X. That is, $D_i \subset X, i = 1, ..., k, and\, U = \{D_1, ..., D_k\}$.

A proximity measure $\wp \; on \; U$ is a function

$$\wp : U * U \longrightarrow R$$

Usually, the proximity measures between two sets Di and Dj are defined in terms of proximity measures between elements of Di and Dj [17].

$$L = \|Y_i - Y_j\| = \sqrt{\sum_{k=1}^{N} (Y_{ik} - Y_{jk})^2} \quad (7)$$

We analyse the dendrogram generated for the hierarchical binary cluster tree using 3 images for each phoneme (ba & ma) and for each speaker, the training images is about 30 images of size 120*160 pixels.We conclude some observations on the intraclass (si) and interclass (se) confusions that were found from this study for the first class of viseme (ba-ma).

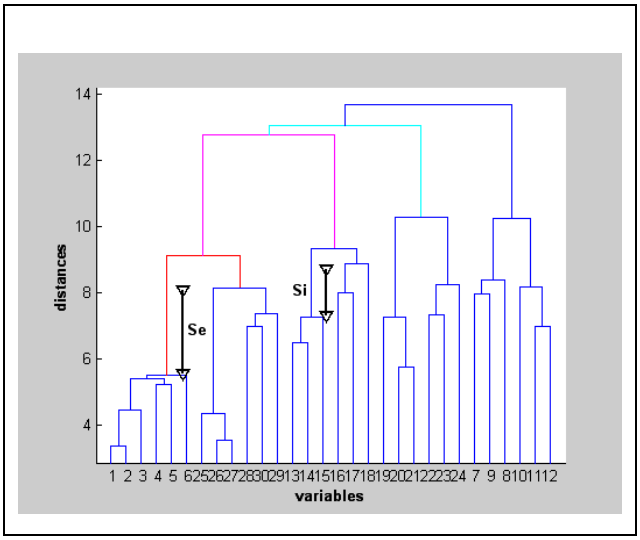

Figure5. Example Hierarchical classification for images (ba &ma)

For the same visemic category, (ba and ma) and also (dja and cha), we investigate the variability of the pitch parameter for the ten speakers, the simulation results are shown in figure 6.

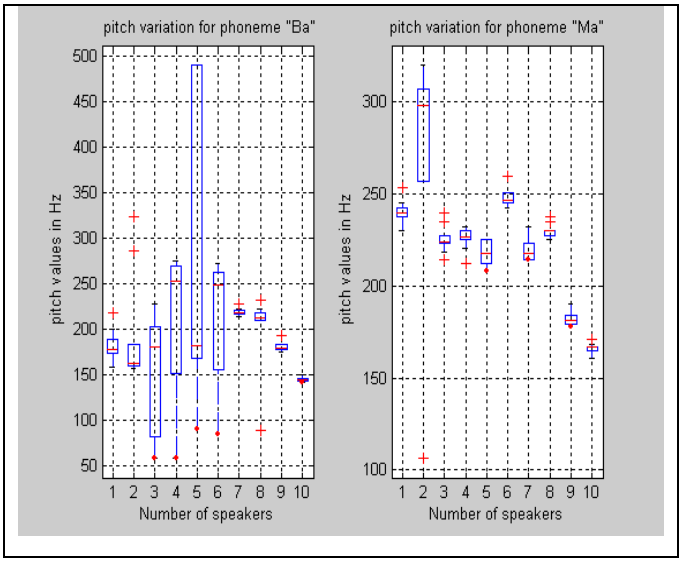

Figure6. Pitch variation for the first viseme

The subplot of the pitch variation demonstrates the intrapersonel and the extrapersonel variation for the same viseme, this study permits us to show the importance of the speech analysis in order to improve the audiovisual speech recognition or the lip reading system for future work.

4.2 MLP Neural Network for viseme classification

Multilayer Perceptron Neural Networks are feed-forward and use the Back-propagation algorithm. We imply feed-forward networks and Back-propagation algorithm (plus full connectivity).

Neural networks are adaptive statistical devices. This means that they can change iteratively the values of their parameters (i.e., the synaptic weights) as a function of their performance. These changes are made according to learning rules which can be characterized as supervised (when a desired output is known and used to compute an error signal) or unsupervised (when no such error signal is used) [18].

Learning process in Backpropagation requires providing pairs of input and target vectors. The output vector y of each input vector is compared with target vector d. In case of difference the weights are adjusted to minimize the difference. Initially random weights and thresholds are assigned to the network.

The network has a three-layered architecture and is trained using the back-propagation algorithm. The number of the input nodes is equal to the size of the input vectors. The number of the output nodes is equal to the number of classes. The number of the hidden nodes is chosen by the user.

Nine neural networks were constructed for each specified viseme. All the NNs trained present fast convergence and the training process terminated within 1000 epochs, with the summed squared error (SSE) reaching the pre-specified goal (10-3) .

Various transfer functions were tested for training the network and average minimum MSE on training (MSEA) is measured; logsigmoid is the most suitable transfer function.

Table 3 gives classification results of the 9 viseme experiments, where the classification rate is defined as the ratio of correctly classified visemes to the total images inputted.

Table 3 . Viseme identification using MLP

Viseme/ symbol	Recognition rate%
1st viseme : Ba	100
2nd viseme : Cha	98
3rd viseme :dha	96
4th viseme :Ha	95
5th viseme :Ta	94
6th viseme :Fa	96
7th viseme :la	100
8th viseme :Ra	100
9th viseme :wa	96

5. CONCLUSION

In this paper we introduced an Arabic viseme system based on visual speech database through statistics analysis. This study demonstrates that the use of statistical parameters like the standard deviation and the geometrical parameters like height and width of lip permits to distinguish visual speech in Standard Arabic.

We've recorded a small visual speech database, All the 28 Arabic consonants were pronounced as monosyllabic sequences, these 11 groups of consonants should probably correspond to the 11 visemes of Arabic, neuronal network

algorithms demonstrate good recognition rate for al visemes studied.

The spatio-temporal characteristics of articulatory movements and their relationships with the co-produced acoustic signal should be taken into consideration for future work.

6. REFERENCES

[1] Waters,keith and Levergood, Thomas M, DECface: An automatic lip-synchronization algorithm for synthetic faces, In Technical report series, CRL 93/4, Digital Equipment Corporation, Cambridge Research Lab, September 23, 1993 .

[2] Breen, Dr. A. P, Bowers Ms. E. and Welsh Dr. W, An Investigation into the generation of mouth shapes for a talking head.

[3] Möttönen, Riikka Olivés Jean-Luc, Kulja Janne and Sams, Mikko, Parameterized visual speech synthesis and its evaluation.

[4] Tiddeman, Bernard and Perret, David, Prototyping and transforming visemes for animated speech,

[5] Abdul Rafay Abbasi and Naveed Ahmad, Urdu Viseme Identification", pp68-71.

[6] Tony Ezzat and Tomaso Poggio, Visual Speech Synthesis by Morphing Visemes, In *A.I. Memo No. 1658 C.B.C, L,* Paper No. 173, Artificial Intelligence ,Laboratory, M.I.T, May 1999.

[7] Michael M. Cohen and Dominic W Massaro, Modeling Coarticulation in synthetic visual speech.

[8] Riikka Möttönen, Jean-Luc Olivés, Janne Kulju, and Mikko Sams, Parameterized visual speech synthesis and its evaluation, Eusipco 2000 X, European Signal Processing Conference,September 4-8, 2000Tampere, Finland.

[9] Hassan Satori, Hussein Hiyassat, Mostafa Harti, and Noureddine Chenfour, Investigation Arabic Speech Recognition Using CMU Sphinx System",The International Arab Journal of Information Technology, Vol. 6, No. 2, April 2009.

[10] Rabiner, L. and B. Juang , Fundamentals of Speech Recognition, Englewood Cliffs, N.J.: Prentice Hall,1993.

[11] WANG Anhong, BAO Huaiqiao, CHEN, Primary research on the viseme system in Standard Chinese,

[12] C. Benoît, T. Lallouache, T. Mohamadi and C. Abry, A set of French visemes for visual speech synthesis, In Talking Machines: Theories Models and Designs, G Bailly and C. Benoît, Editors. Elsevier B.V. p. 485-501.

[13] L.Revéret, Conception et évaluation d'un système de suivi automatique des gestes labiaux en parole, docteur de l'institut national polytechnique de Grenoble,thése préparée au sein de l'institut de la communication parlée.

[14] Salah Werda, Walid Mahdi and Abdelmajid Ben Hamadou, Lip Localization and Viseme Classification for Visual Speech Recognition, International Journal of Computing & Information Sciences Vol.5, No.1, April 2007.

[15] Gerasimos Potamianos, Chalapathy Neti, "Audio-Visual Automatic Speech Recognition": An Overview, Chapter to appear in: Issues in Visual and Audio-Visual Speech Processing, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, Eds., MIT Press, 2004.

[16] Fatma zohra CHELALI , Amar DJERADI, " Primary research on Arabic visemes, Analysis in space and frequency domain", International Journal of Mobile Computing and Multimedia Communication (IJMCMC), published by IGI Global, USA,pp 1-19, DOI: 10.4018/IJMCMC, ISSN: 1937-9412, EISSN: 1937-9404, vol.3 , N°4,2011.

[17] Sergios Theodoridis and Konstantinos koutroumbas. (2003). book *pattern recognition*, second edition, Elsevier (USA).

[18] Herve Abdi,"Neural networks",Program in Cognition and neurosciences,MS:Gr.4.1,The university of Texas at Dallas.

[19] McGurck et J. Mcdonald. "Hearing lips and seeing voice". *Nature*, 264 : 746-748, Decb 1976.

[20] Naotoshi Seo sonots.(2008). Pitch Detection. (report ENEE632 Project4 Part I). March 24, 2008.