# Speech Synthesis from ultrasound and optical images of the speaker's vocal tract

Thomas Hueber

***Abstract*— The goal of my placement at the Electronic Lab of ESPCI [1], in collaboration with LTCI of Télecom Paris, was to investigate the feasibility of a silent speech interface using ultrasound imaging and a profile video of a speaker's head. A method for image preprocessing is proposed, based on the anisotropic diffusion filter. A new algorithm for the extraction of ultrasound image features, called *EigenTongues* is proposed. A method to describe lip profile , based on curvature computation, is introduced. Mel frequency cepstral coefficients are chosen for speech description. Finally, machine learning techniques are used to model the visual-acoustic link between image features and speech signal features.**

## I. Introduction

THERE has been interest recently in the idea of a sensor-based system allowing speech communication via standard articulators, but without glottal activity ; that is, a **silent** speech interface.

The *Ouisper* project[2], on which I worked, proposes to build a device for production of intelligible speech, from ultrasound and optical imagery of the tongue and lips - **without activation of the vocal chords**.

Three parters are involved in this project :

- Laboratoire d'Electronique of ESPCI
- Laboratoire du Traitement et Communication de l'Information (LTCI) of l'ENST, Paris
- Vocal Tract Visualisation Laboratory (VTVL), University of Maryland Dental School

Two distinct types of application can be envisioned :

- an alternative to tracheo-oesophagal speech (TES) for persons having undergone a tracheotomya and a prosthesis for patients who have lost the use of their vocal chords.
- a "silent telephone" for use in situations where quiet must be maintained

The general operation of the *Ouisper* approach is illustrated by the following diagram .

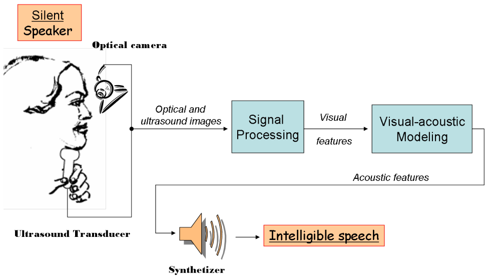


Fig. 1. Operation of the Ouisper system

Currently, the only other system under development for this type of application is one using electromyography [Jorgensen *et al.*, 2003], recently developed by the NASA Ames Lab in the United States.

Promising results using the *Ouisper* technique, have already been published in IEEE conferences by my placement director, Professor Bruce Denby, initiator of the project, in [Denby *et al.*, 2006] and [Denby et Stone, 2004].

The first section briefly presents the architecture, the workings and the main pathology of the voice organ. The second section describes data acquisition. The third section exposes the new methods I proposed to extract features from vocal tract images. Section four is devoted to speech signal description. Finally, machine learning techniques, used to model

[1] Ecole Supérieure de Physique et Chimie Industrielles de la Ville de Paris

[2] Oral Ultrasound Synthetic Speech

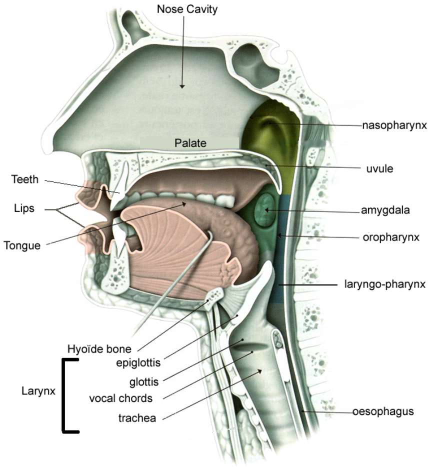


Fig. 2. Voice organ

the visual-acoustic link between visual features and speech signal features, are introduced in the last section.

## II. General presentation of the voice organ

### A. Workings of the voice organ

The voice organ is the part of the human body responsible for the generation of sound, usually in the form of speech or singing. It is comprised of the larynx and the vocal tract. Figure 2 is a sagittal cut of the voice organ.

The human voice produces sounds in the following manner :

- Air pressure from the lungs creates a steady flow of air through the trachea, larynx and pharynx (back of the throat).
- The vocal chords in the larynx vibrate, creating fluctuations in air pressure that are known as sound waves.
- Resonances in the vocal tract modify these waves according to the position and shape of the lips, jaw, tongue, soft palate, and other speech organs, creating formant regions and thus different qualities of sonorant (voiced) sound.
- The mouth and nose openings radiate the sound waves into the environment.

### B. Main voice organ pathology

Because of the medical applications of the Ouisper Project, this section decribes one of the main pathologies of the voice organ, larynx cancer, with focus on its treatment. Age, smoking, alcohol and exposition to dangerous material, are risk factors to get cancer of the larynx. Cancer of the larynx may be treated with a laryngectomy [3] During a laryngectomy, the surgeon may need to make a stoma [4]. The stoma is a new airway through an opening in the front of the neck. Air enters and leaves the trachea and lungs through this opening. After a total laryngectomy, the stoma is permanent. In this case, the patient has to learn to speak in a new way, by using the tracheoesophageal puncture speech. For tracheoesophageal puncture (TEP), the surgeon makes an opening between the trachea and the esophagus. A valve fits into this opening. Patients can cover their stoma with a finger and force air into the esophagus through the valve. Unfortunately, many patients are unable to use this method. They choose to use a mechanical larynx, called electrolarynx. This machine, transmits a vibration noise to the throat which the patient forms into words with his lips, teeth, and tongue. However, speech produced by these different methods is not natural. Patient have to make a big effort to speak. The *Ouisper* project wants to provide an alternative device, able to produce high quality speech, by using ultrasound and optical imagery of the vocal tract. Imagery techniques of the vocal track are described in the next section.

## III. Data acquisition

### A. Ultrasound Principles

Ultrasound imagery is a medical imaging technique that uses high frequency sound waves and their echoes. Ultrasound is an ultra high-frequency sound wave emanating from a piezoelectric crystal[5] that produces an image by using the reflective properties of sound waves. Several crystal elements

[3] surgery to remove part or all of the larynx.

[4] This surgery is called a tracheostomy.

[5] A piezoelectric crystal converts electricity into mechanical vibrations (i.e., sound waves) and vice versa.

are fit in a transducer probe which is the main part of the ultrasound system. The transducer probe makes sound waves which travel into the body and hit a boundary between tissues. Some of the sound waves get reflected back to the probe, while some continue until they reach another boundary and get reflected. The time an echo takes to return to the transducer is converted to distance. The distance and intensity of the echos are displayed as a two dimensional image.

### B. *Voice organ ultrasound imagery*

The Vocal Tract Visualisation Lab (VTVL), involved in the *Ouisper* Project is specialized in ultrasound imaging of the vocal tract. When using ultrasound imagery to visualise the vocal tract, the transducer is placed beneath the chin. During speech, the lips, mouth, chin and jaw, move and specific techniques must be used to maintain the transducer close to the voice organ.

The HATS [6] [Stone, 2003] system developed by M.Stone's team allows the visualisation of the vocal tract during speech. This system is illustrated by the following diagram 3.

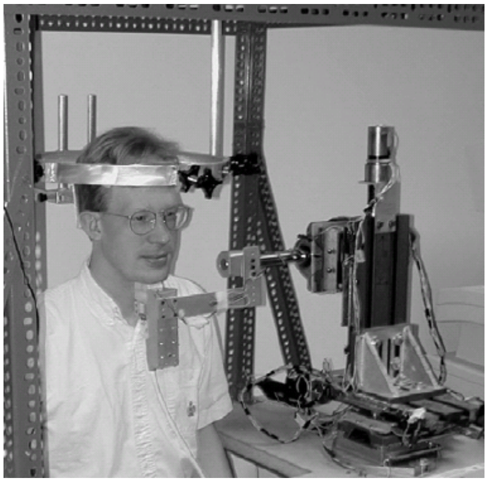


Fig. 3. Head and Transducer Support System - HATS

This specific acquisition system is designed to keep contact between the chin and the transducer, during voice organ motions. This system provides real-time images of the vocal tract during phonation. In addition to ultrasound images, the HATS system provides a profile view of the speaker. A typical image given by VTVL, is shown below in figure 4.

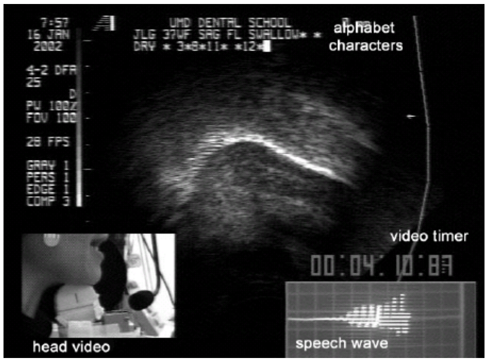


Fig. 4. Ultrasound image of the vocal tract

Different structures can be identified in the ultrasound image :

- The upper tongue surface : The sound wave travels upward through the tongue body until it reaches and reflects back downward from the upper tongue surface.
- The hyoïde bone and mandible bone : These structures refract the sound before it reaches the tongue surface and create an "acoustic shadow" (black region) at both edges of the image.
- The palate bone : the palatal shape should be visible in a frame in which a swallow is occurring.
- Muscle, fat and connective tissue interfaces : The tongue contains considerable amounts of fat, which may refract the sound enough that the returning echo is significantly attenuated.

The HATS system is a powerful tool for speech research since it gives multimodal data of the voice organ during speech. Optical images (a profile view of the speaker's head), ultrasound images and speech are acquired simultaneously. One hour of acquisition has been made by VTVL, for this study. The ability to produce such data allows to consider a modeling of the relation between voice organs

[6] HATS : Head and Transducer Support System

motion and speech signal.

## IV. Optical and ultrasound vocal tract image processing

Relevant information must be extracted from optical and ultrasound images.

### A. Ultrasound images processing

*1) Preprocessing:* To limit useless computation cost, the ultrasound images are reduced to a 50 by 50 grid, superimposed on the original fan-shaped data field and enclosing region of interest, as shown in figure 5.

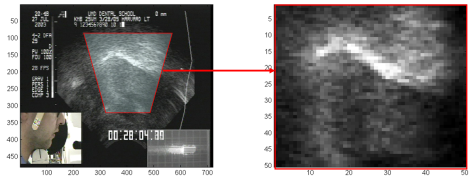


Fig. 5. Ultrasound image reducing

Ultrasound images are characterized by a specific noise, called speckle, which is a multiplicative and locally correlated noise. This noise plagues ultrasound image interpretation and description. For images that contain speckle, an enhancement goal is to remove the speckle without destroying important image features. Speckle filters can be classified in two types. The classic filter, introduced by Lee [Lee, 1980], Kuan [Kuan *et al.*, 1985] and Lopes [Lopès *et al.*, 1990], forms an output image by computing a linear combination of the center pixel intensity in a filter window with the average intensity of the window. So, the filter achieves a balance between straightforward averaging (in homogeneous regions) and the identity filter (where edges and point features exist).This balance depends on the local coefficient of variation inside the moving window defined by :

$$\gamma^2(s) = \frac{1}{|\eta_s|} \sum_{p \in \eta_s} \frac{(I_p - \bar{I}_s)^2}{(\bar{I}_s)^2} \qquad (1)$$

where $\eta_s$ is the filter window centered on $s$ and $\bar{I}_s$ pixel mean intensity on this window.

The second type of speckle adapted filter is the anisotropic diffusion filter, introduced by [Perona et Malik, 1990]. Perona & Malik formulate the anisotropic diffusion filter as a diffusion process that encourages intraregion smoothing while inhibiting interregion smoothing. Mathematically, the process is defined as follows :

$$\frac{\partial I}{\partial t} = div(c(|\nabla I|)\nabla I) \qquad (2)$$

where $\nabla$ is the gradient operator,*div* the divergence operator, $||$ denotes the magnitude, $t$ refers to the iteration step $c(x)$ the diffusion function which is a monotonically decreasing function of the image gradient magnitude:

$$g(s) = \frac{1}{1 + (\lambda s)^2} \qquad (3)$$

where $\lambda$ is an edge magnitude parameter. In the anisotropic diffusion method, the gradient magnitude is used to detect an image edge or boundary as a step discontinuity in intensity. If $|\nabla u| \gg \lambda$ then $g(|\nabla u|) \to 0$ and we have an all-pass filter; if $|\nabla u| \ll \lambda$ then $g(|\nabla u|) \to 1$ and we achieve isotropic diffusion. The advantages of anisotropic diffusion include intra-region smoothing and edge preservation.

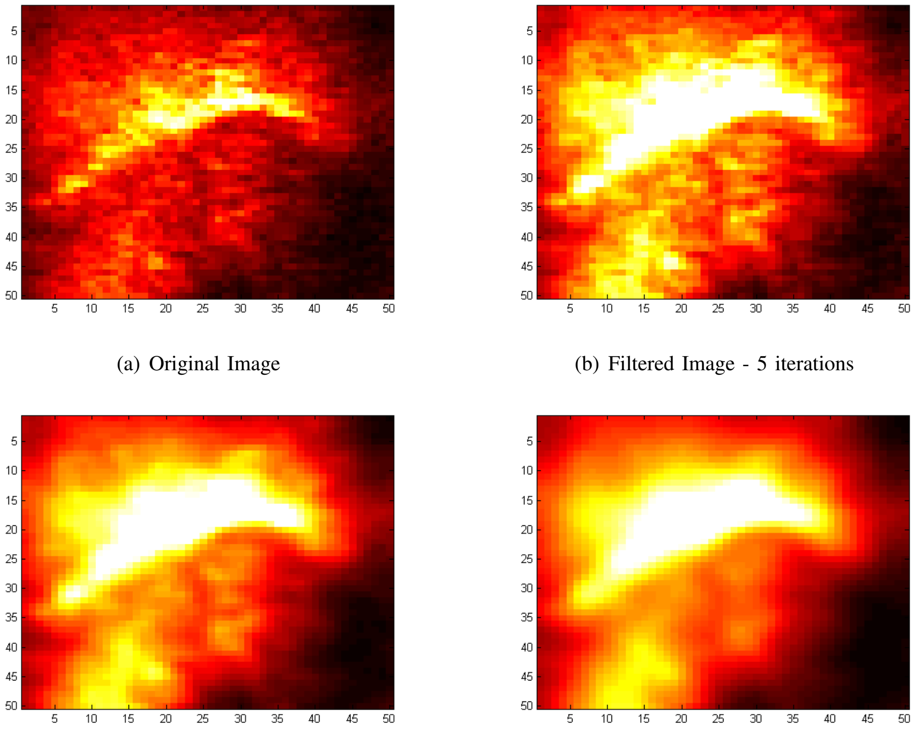
In [Yu et Acton, 2002], Yu combined the Lee and Perona & Malik approach by proposing a new anisotropic diffusion method for smoothing speckled imagery. Yu merged equations 4 and 2 in a new partial differential equation :

$$\frac{\partial I}{\partial t} = \frac{1}{|\bar{\eta}_s|} div[c(\gamma_t)\nabla I] \qquad (4)$$

where $\gamma_t$ can be called an "instantaneous coefficient of variation".

This filter has been implemented and applied on reduced ultrasound images of the vocal tract. Results are illustrated by figure 6.

*2) The* EigenTongues *approach:* In [M.Li *et al.*, 2003], Li presents *Edge Track*, a program for tongue edge extraction from ultrasound images of the vocal tract. In [Denby *et al.*, 2006], Denby uses a similar algorithm to extract tongue edge in each frame of the data set. Denby considers that the most relevant information in ultrasound



(a) Original Image

(b) Filtered Image - 5 iterations

(c) Filtered Image - 20 iterations

(d) Filtered Image - 40 iterations

Fig. 6. Results of the Anisotropic Diffusion Filter

frame sequence is tongue motion. However, other structures seem to move during speech, and their position could be important for visual-acoustic modeling. That's why, a new approach, based on a global image description, has been proposed. This method is inspired from Turk's work, described in [Turk et Pentland, 1991].

Let a reduced image $I(x, y)$ be a two-dimensional $N$ by $N$ array of intensity values or a vector of dimension $N^2$. A reduced vocal tract image describes a vector of dimension 2500, or equivalently, a point in a 2500-dimensional space. A ensemble of images is a collection of points in this huge space. Filtering has reduced random behavior between closed images in the video sequence. Thus, all frames of the sequence are quite similar. They will not be randomly distributed in this huge space, and can be described by a relatively low-dimensional subspace. The main idea of the method is to use a principle components analysis to find the vector which best accounts for the distribution of the vocal tract image, within the entire image space. These principle components are vectors which define the subspace of the vocal tract image. We call this space the *TongueSpace*. Each vector of length $N^2 = 2500$, describes a 50 by 50 image, and is a linear combination of the original vocal tract images. We call them *EigenTongues*, as Turk's vectors are called *EigenFaces*. Some examples of *EigenTongues* are shown in figure 7.

We show that a new image, which has not participated in the *TongueSpace* building can be encoded by a small number of its first coordinates in this space. Figure 8 illustrates an original frame and its

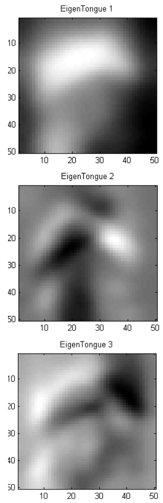


Fig. 7. *EigenTongues*

re-synthesis by using only its 40 first coordinates in the *TongueSpace*.

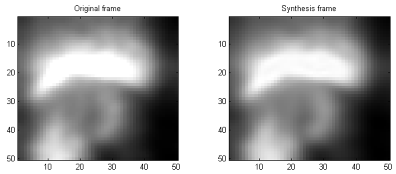


Fig. 8. Description of a new image with 40 *EigenTongues*

The 40 first coordinates of an image in the *TongueSpace* seem to be relevant visual features.

## B. *Optical image processing*

Edge extraction of the lips profile is simply done by a Sobel method. In [Denby *et al.*, 2006], by working on a 90 degrees rotated frame, Denby can consider the lips edge as a function, and find the lower and upper lips by searching its maxima. However, this method is not speaker independant, as show in figure 9. Because a lips profile can not always be considered as a function, a new approach has been proposed. In [Feldman et Singh, 2005], Feldman used the *Turning angle* to easily compute the curvature of a two-dimensional curve. The approach is based on Attneave's idea, that information along a visual contour is concentrated in regions of high magnitude of curvature, rather than distributed uniformly [Attneave, 1954]. The lower and upper lips belong to high curvature regions, and the *turning angle* curve gives speaker-independant and relevant features for lips profile description, as shown in figure 10.

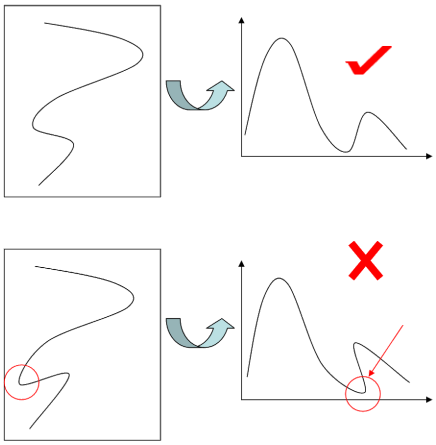


Fig. 9. Lips profile particularity

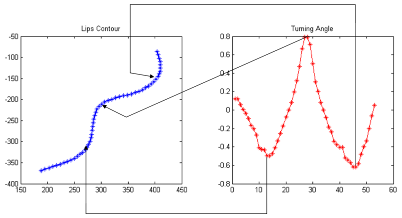


Fig. 10. Lips profile description thanks to turning angle computation

Visual features have been extracted from ultrasound and optical images of (part of) the voice organ.

# V. Features extraction from speech signal

In [Denby *et al.*, 2006], Denby use the Line Spectrum Frequencies (LSF) [Zheng *et al.*, 1998] representation of the speech signal. High quality speech synthesis can be achieve only with a good fundamental frequency estimation. However, the *Ouisper* device is a **silent** speech communication interface, and requires no vocalisation, no glottis activity, and thus no information on fundamental frequency. The LTCI laboratory is the originator of the ALISP system (Automatic Language Independent Speech Processing) [Cernocky, 1998], [Mosbah, 2005] which allows high quality speech synthesis based on a statistical segmentation of a large speech corpus [Bimbot *et al.*, ], and a concatenative high-quality re-synthesis of speech. In order to embed the ALISP system in the *Ouisper* project, Mel Frequency Cepstral Coefficients [Davis et Mermelstein, 1980] are introduced here.

Many experiments have shown that the ear's perception of the frequency components in the speech does not follow the linear scale but the mel-frequency scale, which should be understood as a linear frequency spacing below 1kHz and logarithmic spacing above 1kHz. The common used formula to approximately reflex the relation between the mel-frequency and the physical frequency is given by,

$$M = \frac{1000}{\log_2} * \log(1 + \frac{f}{1000}) \qquad (5)$$

where $f$ is frequency in Hz.

The system diagram to compute the MFCC and classification scheme is shown in Figure 11 and is briefly explained below.

A segment of speech is hamming windowed and transformed to the frequency domain via the fast Fourier transform, and then the magnitude spectrum of the utterance is passed through a bank of triangular shaped filters whose center frequencies are spaced along the perceptually motivated mel frequency scale. The energy output from each filter is then log-compressed and transformed to the cepstral domain via the DCT.

MFCC's coefficients are known to be relevant and robust features of the speech signal.

## VI. Visual-acoustic Modeling

Visual and acoustical features are extracted from video and audio sequences of the data set provided by VTVL. A vector $x$ of 50 visual features describes each frame and a vector $y$ of 12 MFCC coefficients simultaneously describes the correspondent speech signal. A neural network, called multi layer perceptron (MLP) is used to perform the mapping between the 50 input variables and the 12 MFCC's.

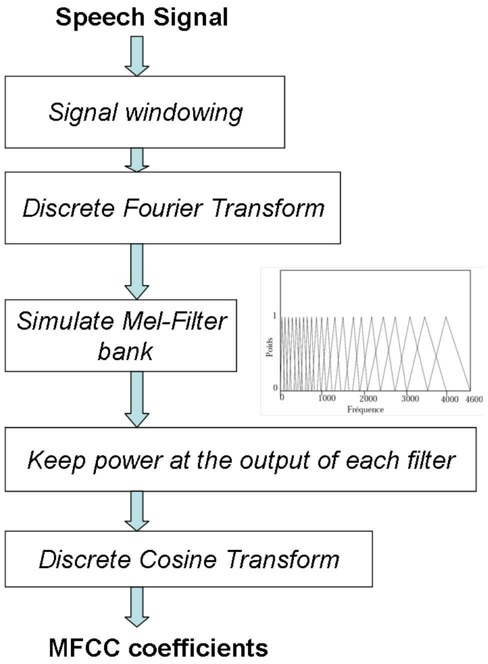


Fig. 11. Mel Frequency Cepstral Coefficients computation diagram

### A. Non-linear modeling with neural networks

Neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs. A neural network is composed of a large number of highly interconnected processing elements, called artificial neurones, working together to solve specific problems, such as **non-linear regression problems**.

Artificial neurones are functions from many dimensions to one dimension. They receive one or more inputs and sum them to produce an output. This output is passed through a non-linear function called activation function. A graphical representation of an artificial neurone is shown in figure 12.

MLP consists of multiple layers of artificial neurones, usually interconnected in a feed-forward way. Each neuron in one layer has directed connections to the neurons of the subsequent layer, as shown in figure 13.

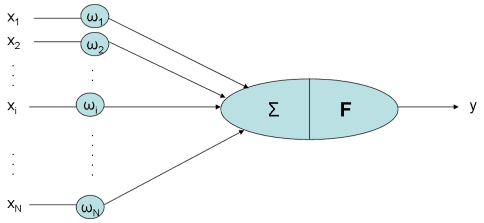


Fig. 12. Graphical representation of an artificial neurone

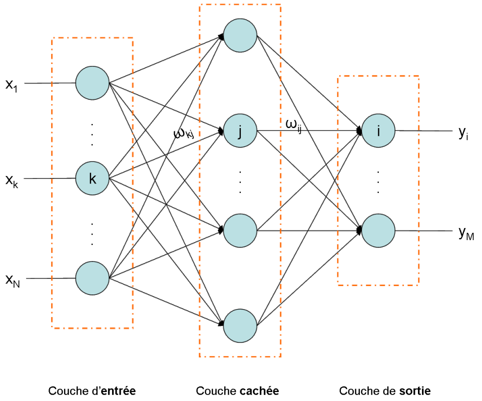


Fig. 13. Multi Layer Perceptron

During the supervised training, a learning set of example (x,y) pairs is given to the network which adjusts its parameters to find the function $f$ such as $y = f(x)$. In the test, MLP is normally able to predict output from new examples. The reader is invited to consult [Dreyfus *et al.*, 2004] for more information about neural networks.

For visual-acoustic modeling with neural MLP, we use in this study, 71 595 learning examples and 878 test examples. For MFCC's prediction, we use 12 networks with 20 hidden neurones and 50 inputs.

### B. Results

Two methods can be used to evaluate the quality of the modeling. The first one is the computation of the mean square error between model predictions and true values on the test database, called $\alpha$. The second one is a scatter plot "Predicted value *vs.* Original value". Both methods are used to describe the result of this study. Table I presents $alpha$ values, for each acoustic feature (*i.e* each MFCC).

| . | $\alpha$ | . | $\alpha$ |
|---|---|---|---|
| $MFCC_1$ | 0.17 | $MFCC_2$ | 0.13 |
| $MFCC_3$ | 0.14 | $MFCC_4$ | 0.11 |
| $MFCC_5$ | 0.12 | $MFCC_6$ | 0.11 |
| $MFCC_7$ | 0.12 | $MFCC_8$ | 0.13 |
| $MFCC_9$ | 0.13 | $MFCC_{10}$ | 0.11 |
| $MFCC_{11}$ | 0.10 | $MFCC_{12}$ | 0.10 |

TABLE I
MFCC COEFFICIENTS PREDICTION USING A MULTI LAYER PERCEPTRON

Scatter plots are displayed in Figure 14.

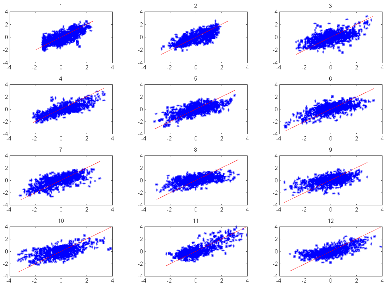


Fig. 14. Scatter plot of test results for the MFCC's. Horyzontal axis : True MFCC ; vertically : Predicted MFCC's

MFCC's 4,5,6,7-11 appear to be the easiest to learn from the image of the vocal tract. The model we propose is able to predict acoustical description of speech with a 10 % to 20 % error.

## VII. CONCLUSION

In [Denby *et al.*, 2006], Denby presents a modeling of LSF coefficient from tongue contours and lip profile features. This model is not sufficient for learning the LSF of silent and fricative speech frames. In this study, it has been shown that MFCC coefficients can be predicted with new lips profile features and the *EigenTongues* new approach, with no distinction on type of speech frame (voiced or unvoiced). The ALISP system will soon be used for speech synthesis from predicted MFCC coefficients sequences. This will be one of my first jobs during my PhD on the *Ouisper* project.

## Acknowledgment

I'd like to thank all the members of ESPCI Electronic Lab, especially its director Gérard Dreyfus, who accepted me in his team to do this work,

Gérard Chollet, research director at Télécom Paris LTCI Lab , codirector of my placement, for his precious remarks and encouragements,

Isabelle Block and Guido Aversano (LTCI, Télecom Paris), for their time and expertise.

I'd like to express my deepest gratitude to Bruce Denby, researcher at ESPCI and Professor at Université Pierre et Marie Curie, director of my placement, for his trust and his availability.

## References

[Attneave, 1954] Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61:183–193.

[Bimbot *et al.*, ] Bimbot, F., Chollet, G., P.Deleglise et Montacie, C. Temporal decomposition and acoustic-phonetic decoding of speech. *International Conference on Acoustics, Speech, and Signal Processing*.

[Cernocky, 1998] Cernocky, J. (1998). *Speech processing using automatically derived segmental units : applications to very low bitrate coding and speaker verification*. Thèse de doctorat, Université Paris VI.

[Davis et Mermelstein, 1980] Davis, S. et Mermelstein, P. (1980). Comparison of parametric representation for monon-syllabic word recognition in continuously spoken sentences. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 28(4):357–366.

[Denby *et al.*, 2006] Denby, B., Oussar, Y., Dreyfus, G. et Stone, M. (2006). Prospect for a silent speech interface using ultrasound imaging. *International Conference on Communication Audio and Speech Processing*.

[Denby et Stone, 2004] Denby, B. et Stone, M. (2004). Speech synthesis from real time ultrasound images of the tongue. *International Conference on Communication Audio and Speech Processing*.

[Dreyfus *et al.*, 2004] Dreyfus, G., Samuelides, M., Martinez, J., Gordon, M., Badran, F., Thiria, S. et Hérault, L. (2004). Réseaux de neurones, ed eyrolles, collection algorithmes.

[Feldman et Singh, 2005] Feldman, J. et Singh, M. (2005). Information along contours and object boundaries. *Psychol Rev*, 112(1):243–252.

[Jorgensen *et al.*, 2003] Jorgensen, C., Lee, D. et Agabon, S. (2003). Sub auditory speech recognition based on emg/epg signals. *Proceedings of the International Joint Conference on Neural Networks*, 4,:3128–3133.

[Kuan *et al.*, 1985] Kuan, D., Sawchuk, A., Strand, T. et Chavel, P. (1985). Adaptative noise smoothing filter for images with signal dependant noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-7, n°2, p.165-177*.

[Lee, 1980] Lee, J. (1980). Digital image enhancement and noise filtering by use of local statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-2, p. 165-168*.

[Lopès *et al.*, 1990] Lopès, A., Touzi, R. et Nezry, E. (1990). Adaptative speckle filters and scene heterogeneity. *IEEE Transactions on Geoscience and Remote Sensing, vol. 28, n°6, p. 992-1000*.

[M.Li *et al.*, 2003] M.Li, C.Kambhamettu et M.Stone (2003). Edgetrak, a program for band-edge extraction and its applications. *Sixth IASTED International Conference on Computers Graphics and Imaging*.

[Mosbah, 2005] Mosbah, B. B. (2005). *Utilisation de la mémoire de parole pour la reconnaissance (Application pour des personnes handicapées)*. Thèse de doctorat, Télécom Paris.

[Perona et Malik, 1990] Perona, P. et Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(7):629–639.

[Stone, 2003] Stone, M. (2003). A guide to analysing tongue motion from ultrasound images. *Clinical linguistics and phonetics*, pages 359–366.

[Turk et Pentland, 1991] Turk, M. A. et Pentland, A. P. (1991). Face recognition using eigenfaces. *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on*, pages 586–591.

[Yu et Acton, 2002] Yu, Y. et Acton, S. T. (2002). Speckle reducing anisotropic diffusion. *IEEE Transactions on Image Processing*, 11(11):1260–1270.

[Zheng *et al.*, 1998] Zheng, F., Song, Z., Li, L., Yu, W., Zheng, F. et Wu, W. (1998). The distance measure for line spectrum pairs applied to speech recognition. *In Int. Conf. on Spoken Language Processing (ICSLP-98)*, pages 3:1123–1126.