Control Prosody using Multi-Agent System

Kenji Matsui[a], Kenta Kimura[a], Alberto Pérez[b]

[a] Osaka Institute of Technology, Osaka, Japan

[b] Computer and Automation Department, University of Salamanca, Spain

KEYWORD

Prosody
Electrolarynx
Hands-free
Multi-agent system
Agents

ABSTRACT

Persons who have undergone a laryngectomy have a few options to partially restore speech but no completely satisfactory device. Even though the use of an electrolarynx (EL) is the easiest way for a patient to produce speech, it does not produce a natural tone and appearance is far from normal. Because of that and the fact that none of them are hands-free, the feasibility of using a motion sensor to replace a conventional EL user interface has been explored. A mobile device motion sensor with multi-agent platform has been used to investigate on/off and pitch frequency control capability. A very small battery operated ARM-based control unit has also been developed to evaluate the motion sensor based user-interface. This control unit is placed on the wrist and the vibration device against the throat using support bandage. Two different conversion methods were used for the forearm tilt angle to pitch frequency conversion: linear mapping method and F0 template-based method A perceptual evaluation has been performed with two well-trained normal speakers and ten subjects. The results of the evaluation study showed that both methods are able to produce better speech quality in terms of the naturalness.

1 Introduction and background

Due to technological advances, intelligent systems – providing a huge range of facilities – have become an important element in our daily lives. Thanks to these new intelligent systems, disabled people are able to overcome difficulties. Patients who have undergone a laryngectomy have several options for the restoration of speech, but no currently available device is satisfactory. An EL introduces a source of vibration in the vocal tract by producing vibrations in the external wall. Even though this is the easiest way for patients to master, the intelligibility of consonants is diminished and the speech is uttered at a monotone frequency since it does not produce airflow. As an alternative, the esophageal speech does not require any special equipment, but does require the speaker to insufflate air into the esophagus and it limits the pitch range and intensity. Both esophageal and tracheo-esophageal speech are characterized by low average pitch frequency large cycle-to-cycle perturbation in pitch frequencies, and low average intensity. As for utilizing esophageal speech, it has been found that age is a really important factor. As patients get older, they face difficulty in mastering the esophageal speech. Because of that, the EL is an important device even for the people using esophageal speech.

Regarding the advantages of the EL, to begin, a patient is able to speak using long and easily understandable sentences. Additionally, no special care requirements are needed; The EL must simply be turned on and placed against the neck. Finally, the EL can be used by almost everybody, regardless of the post-operative changes in the neck. In those few cases where scarring prevents proper placement of the EL, an intraoral version can be used.

On the other hand, there are also a couple of disadvantages. Firstly, the EL has a very mechanical tone that does not sound natural. There is usually little change in pitch or modulation. Secondly, the appearance of the EL is far from normal. Pitch frequency control is one of the important mechanisms for EL users to be able to generate naturally sounding speech. There are some commercially available EL devices using a single push button with a pressure sensor to produce F0-contours [SECOM, 2014], [Griffin laboratories, 2014].

There are also similar studies of pitch controlling methods [Kikuchi, Y. et al., 2004], [Takahashi, H. et al., 2001]. However, none of them are hands-free. Some approaches for generating F0-contour without manual interaction have been proposed. Saikachi et al. use the amplitude variation of EL speech [Saikachi, Y. et al., 2009]. Another approach is to generate an F0-contour using an air-pressure sensor that is put on the stoma [Uemi, N. et al., 1994], [Nakamura, K. et al., 2010]. Also, recently, a machine learning F0-contour generation from the EL speech has been proposed [Fuchs, A. K. et al., 2012]. Although most of those studies are in the early stage of research, the results are already showing substantial improvement of EL speech quality. An EL system that has a hands-free user interface could be useful to enhance communication by alaryngeal talkers. Also, the appearance can be almost normal due to the fact that users do not need to hold the transducer against the neck. Practically everyone gestures when speaking, so it would be quite convenient if EL users could use gestures to control the device. Furthermore, gesture control has a lot of potential to handle not only just on/off function, but many other functions as well because hands can generate various types of motion. However, if users are not even able to use hands to gesture, for example, controlling something, we need to consider other part of body movement, or a completely different technique, such as EMG based hands-free EL control [Kubert, H. L. et al., 2009].

As for the total system management issue, multi-agent technology is one of the key technology trends. There are many multi-agent frameworks to help and facilitate working with agents [Poslad, S. et al., 2000], [Argente, E. et al., 2004], [McCabe, K. L. et al., 1995], Bordini, J. F. et al., 2005]. The main drawback of these systems is that they are general purposes. General purpose was considered as a significant issue twenty years ago, but it is not today, on a time when personal computers and mobile phones devices alike have grown exponentially. In addition, the needed architecture must be able to assume the necessary tasks for the integration of disabled people. Moreover, differences among people with disabilities are very different from one to another. Some of the most known European multi-agent systems projects oriented in the direction of our research direction are [CommonWell Project, 2010], [Monami project, 2010], [DISCATEL, 2010], [INREDIS, 2011], [INCLUTEC, 2011].

This study has been undertaken to explore the feasibility of using multi-agent technology and mobile devices (replacing the conventional EL user interface) to control both on/off function, and pitch frequency. The specific goals were: 1) to make a natural generated speech, and 2) to make sure the appearance of the EL against the neck is as natural as possible.

This paper is structured as follows: section 2 presents a revision on the system requirements as indicated by the participants involved in the study. Afterwards, the system implementation is approached for the reader to fully understand the system design. Finally the end of the article briefly explains the results and conclusions obtained from the study.

2 System Requirements

A set of techniques - including user observations, interviews, and questionnaires - were used to understand implicit user needs. The total number of laryngectomized participants in the questionnaire survey was 121 (87% male, 13% female), including 65% esophageal talkers, 12% EL users, 7% both and 21% using writing messages to communicate.

Almost all of the participants claimed that most public areas are difficult for oral communication due to the noisy environment. Typical public areas include train stations, inside of train cars, inside of vehicles, restaurants/pubs, and conventions/gatherings.

The noisy environment issue is a well-known problem and people usually use portable amplifier; however, we have been investigating a smaller, lighter, and low profile speech enhancement system for both esophageal speech [Matsui, K. et al., 2002] and EL. Other needs confirmed from the survey are: (i) natural sounding voice, without a mechanical tone. (ii) Light weight device. (iii) Smaller device, low profile. (iv) Hands-free and easy to use. (v) Low cost. Based on those survey results, the present study was conducted to meet the essential user needs.

3 System Implementation

The system implementation was carried out by using PANGEA (Platform for Automatic coNstruction of orGanizations of intElligent Agents) [Zato, C. et al., 2012a], [Zato, C. et al., 2012b]. It is modeled as a virtual organization of agents. These agents are connected to the PANGEA architecture, a multi-agent architecture designed on the basis of virtual organizations, aimed at creating intelligent environments which are able to be connected to any kind of device, as explained in [Zato, C. et al., 2012c]. The system uses a smartphone that has powerful processing capabilities, provides any functionality needed to connect to PANGEA, and allows the use of its integrated accelerometer to calculate the desired output.
The following subsections explains how the system is integrated, paying special attention to the UI design, the integration of the system with PANGEA platform – multi-agent design, the design of the hardware to include in the system, and the two algorithms used to contrast data.

3.1. Hands-Free UI Design: Gesture and Pitch Control

Gesture control UI can be developed through the use of a system based on photo detector, camera, or accelerometer. Based on the survey results, a three-axis MEMS accelerometer was used in this study. MEMS sensors are very small, low cost, and fit the system requirements well [Matsui, K., et al., 2013].
A MEMS accelerometer accurately measures acceleration, tilt, shock and vibration in applications. The challenge in designing the pitch control algorithm that uses a MEMS accelerometer output to control pitch contour is to reconcile the numerical ranges between two types of data. MEMS output bytes are integers within the range -128 to 127 for a range of ±2G. This issue can often be easily reconciled by linear mapping of one range of values (such as MEMS data values - 128 to 127) into another range (such as 67 to 205 expected as the typical male pitch range).
Another possible pitch control method is to utilize a pitch contour generation model, such as Fujisaki's model [Fujisaki, H. et al., 1988]. The system needs to have a strategy to generate both the phrase component and the accent component from the MEMS output. The F0 template-based method is easier to generate relatively stable pitch contour, however, it may lose some flexibility to generate various pitch patterns.
In this study, both the simple linear mapping method and the F0 template-based method were prototyped and examined to evaluate pitch control performance. Also, the comparison study was performed between conventional EL, the linear map-ping method and the F0 template-based method.

```
<test> help

<sensorAgent> C000
<sensorAgent> C001: getValue <int: number_of_values>: This message
  returns the specified number of last read values, from 1 to 10.
<sensorAgent> C002 End of help
```

Figure 1: Message format offered by the agent sensor in PANGEA

3.2. Multi-Agent Design

In order to integrate the system with the PANGEA platform, a virtual organization was developed, named the Alaryngeal Talkers Organization. This organization includes the following three kinds of agents, all of which improve the complete system because of the inherent advantages to a multi-agent structure.

- *Sensor agent:* this kind of agent is in charge of obtaining measures from the smartphone's accelerometer sensor and providing this data when required by other agent members of the system who are authorized to communicate with it.
- *Config agent*: this kind of agent allows establishing certain configuration data, which are required when establishing a pitch frequency to be the base when re-adjusting the frequency. This is an important factor to fit the frequency with a person's physical appearance, which will be estimated from the entered data. With this, an even more natural result is achieved.

```
<test> help
<configAgent> C000
<configAgent> C001: setGender <bool: gender>: This message sets the
  gender of the current user. 0=male, 1=female.
<configAgent> C001: setWeight <int: kilograms>: This message sets the
  weight of the current user.
<configAgent> C001: setHeight <int: centimeters>: This message sets
  the height of the current user.
<configAgent> C001: bool getGender: This message returns the gender
  of the current user if it is established.
<configAgent> C001: int getWeight: This message returns the weight of
  the current user in kilograms
<configAgent> C001: int getHight: This message returns the height of
  the current user in centimeters.
<configAgent> C002 End of help
```

Figure 2: M Message format offered by the configuration agent in PANGEA

- *Analogic agent*: this kind of agent is responsible for generating and providing an analogue output from the data obtained from the agents involved (sensor and configuration agents). These agents can now only communicate with each other and with control agents that the PANGEA architecture offers, but with the possibility of eliminating this restriction or even expanding the system in future extensions.

3.3. Hardware System Design

We have been using Android-based mobile devices. Android is an open-source operating system (OS) and has a large market share in terms of OS for smartphones and PC tablets. The basic idea is to utilize an accelerometer of an Android mobile device to control on/off function and pitch frequency. Users can control without seeing the display using sensors.

A block diagram of the hardware architecture is shown in Figure **4.** An Android mobile device sends PWM signals to a pair of EL transducers through an amplifier. The Amplifier requires a 9V battery so that the EL transducers can generate sufficient speech output.

Figure 3: Amplifier in a box

The EL transducer is placed against the neck with the neck-bandage. Figure **6** shows the entire system, including the EL transducer and the amplifier.

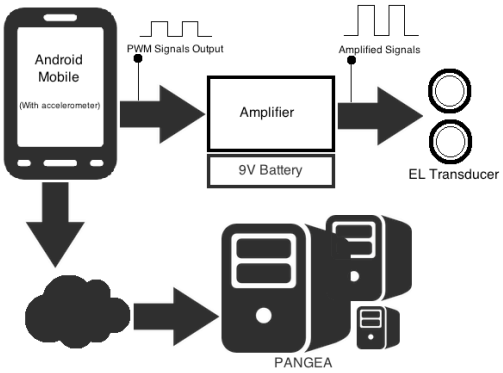

Figure 4: Block diagram of the Hardware Architecture

Figure 5: A pair of EL Transducer

Figure 6: Complete system

3.4. Pitch Control (linear mapping method)

Hand gestures are a very important part of language. A preliminary UI study using forearm movement was conducted in order to evaluate the feasibility of the pitch control mechanism. Figure **7** shows the forearm tilt and the MEMS output (x-axis) when the controller was placed on the wrist. The normal pitch control zone

extends from the horizontal position (0°) to the 75° upward position. The fading out zone extends from the horizontal position to the -25° downward position, and is where the phrase ending pitch pattern is adjusted based on the forearm moving speed. As for the conversion from the MEMS output to the pitch frequency, there are four pitch ranges. Figure **8** shows the relationship between the MEMS output and the four ranges of pitch frequency, i.e. high, mid-high, mid-low, and low. Users can select one of the four ranges.

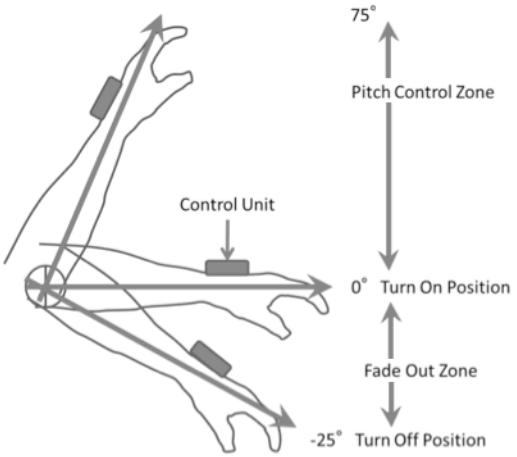

Figure 7: Forearm tilt and pitch control

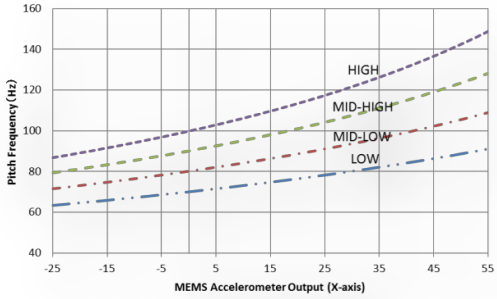

Figure 8: Relation between MEMS Output and Pitch (linear mapping method)

3.5. Pitch Control (F0 template-based method)

The linear mapping method is a straightforward approach; however, it requires precise sensor control to avoid unnatural pitch behavior. The F0 template-based method applies a basic F0 template to the fine F0 contour generation. The phrase component of Fujisaki's model was used to generate the F0 template F0(t). While the system is intended to generate both phrase control and accent control, during the first step of testing the template, we utilized only the phrase component.

$$\ln F0(t) = \ln F\,\mathrm{min} + Ap \cdot Gp(t) \quad (1)$$

where

$$Gp(t) = \alpha^2 t \exp(-\alpha t) \quad (2)$$

The symbols in equations (1) and (2) indicate:

- *F*min is the minimum value of the speaker's F0.
- *Ap* is the magnitude of phrase command.
- α is natural angular frequency of the phrase control mechanism.

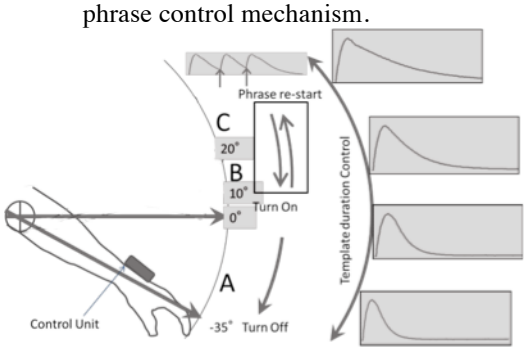

Figure 9: Relation between Forearm tilt and F0 template generation (F0 model-based Method) A-zone: -35°~0°, B-zone: 0°~20° and C-zone: 20°~.

In this study, those values are: *F*min = 80Hz, α=1.5 and *Ap* = 0.75. The calculated F0 template data is stored in the controller software. Figure **9** shows the F0 contour generation mechanism using the MEMS sensor output and the F0 template. The oscillation starts at 10° upward from the horizontal position. The template duration is controlled based on the forearm tilt angle as shown in Figure **9**. The figure also shows how to re-start the F0 template. Basically, the forearm movement (C-zone -> B-zone -> C-zone) is required. A-zone is -35°~0°, B-zone is 0°~20°, and C-zone is 20°~, respectively.

4 Evaluation and Results

Although we confirmed the two types of pitch control functions using the mobile device based

system, this time, we took an ARM-based hardware unit for the pitch control algorithm evaluation. Subjective evaluation tests (by rating scale method) were made with 2 male well-trained normal speakers, and 10 (one female and nine male) subjects. Each speaker read the phonetically balanced test materials as shown in table 1. We used one commercially available EL device (SECOM EL- X0010), prototype-A (linear mapping method, with 70Hz mode), and prototype-B (F0 template-based method). Those 60 speech stimuli (2 speakers ∗3 devices ∗10 sentences) were recorded, and two sets of differently randomized stimuli were prepared. 5 subjects evaluated one set of stimuli, and another 5 subjects rated the other set of stimuli. Each speech stimulus was presented twice.

1	Papa mo mama mo minna de mamemaki o shita.
2	takai takai tokoro e nobotte iku tokoro da.
3	Achirakara mo kochirakara mo dochirakara mo ikukoto ga dekiru.
4	Aoi o ueru.
5	Anohito wa bunkajin to yobareru no ga fusawashî.
6	Shichi gatsu kara hanshin densha de tûkin shite ima su.
7	Ginkô mo gakkô mo aruite ikeru kyori ni ari masu.
8	Kinkô ga tore te iru no de kakkô ga yoi.
9	shijû gamu o kamuno ga syûkan ni natte iru.
10	Hana o ottari ana o hottari sanzanna meni atta.

Table 1: Phonetically balanced Japanese test sentences

The speech stimuli of the subjects were rated in terms of “intelligibility (Clarity)”, “naturalness of the prosody”, and “stability of the prosody” using five level scaling. The subjective evaluation indicated that both prototype-A(LM) and B(FU) obtained higher naturalness scores than the EL de-vice(EL). On the other hand, intelligibility (clarity) and stability shows almost no difference among those devices.

	LM	**FU**	**EL**
Intelligibility	2.99	2.98	2.98
Naturalness	3.11	3.03	2.255
Stability	3.3	3.17	3.345

Table 2: Average evaluation scores

Without losing intelligibility (clarity) and stability of the prosody, both prototype-A and B showed substantial improvement in terms of the naturalness of the prosody. Results of this study indicate that both, usability and speech quality of EL speakers, could be improved by MEMS accelerometer based hands-free UI controller. The ability to control the pitch contour of EL speech with the proposed linear mapping method and F0 template-based method implies that hand gesture control may be adequate for implementation of the hands free user interface for the EL device. Our assumption about the performance difference between the two proposed methods is that the F0 template-based method may be easier to learn and the pitch contour easier to stabilize. However, there was almost no difference between those two methods. We plan to run the same evaluation with actual EL-users, and confirm if the proposed methods perform similarly. Also, a more detailed and precise study across the talkers, sentences, and learning curve has to be performed. As for the gesture control, we tested only the forearm movement; however, it is necessary to test other body locations where users might be able to control the EL device more easily and naturally. According to the user requirements, the evaluation of appearance also needs to be considered. In the study, we set a relatively narrow pitch range in order to avoid wild swings in pitch. A better pitch control range needs to be investigated.

5 Conclusions

An MEMS accelerometer, integrated in a smartphone, hands free UI for EL device was proposed. A hand gesture system was designed and prototyped using a smartphone. Two types of pitch contour generation methods were proposed and tested together with conventional EL device. Results of the evaluation indicated that the proposed methods have a potential to make the EL output prosody more natural, easy to use, and with a less distinct appearance. In addition, the developed multi-agent system provides several advantages. A simple application with multi-profile capacity is achieved, which allows the speaker to obtain an even more natural way of speech. Similarly, the system could be expanded in terms of sensors or even complexity thanks to the characteristics provided by the integration with PANGEA platform [Sánchez, A. et al., 2013], the multi-agent architecture used. It also allows us to keep a record of all messages produced in the system,

which can lead to future studies to improve the system based on the generated knowledge.
A disadvantage of using the PANGEA platform is that the mobile device must necessarily have a connection with the server, either by local network or the Internet. For a situation where a connection is not possible, there is an alternative design, already developed and presented, which can be seen in [Matsui, K. et al., 2013].
However, a more detailed and precise study across the talkers, sentences, and learning curve has to be performed.

6 References

[SECOM, 2013] SECOM company Ltd., Electrolarynx "MY VOICE", [http://www.secom.co.jp/personal/medical/myvoice.html]. Accessed in October 2013.

[Griffin laboratories, 2013] Griffin laboratories. Instruction Manuals, [http://www.griffinlab.com/Help.html]. Accessed in November 2013.

[Kikuchi, Y. et al., 2004] Y. Kikuchi, and H. Kasuya: "Development and evaluation of pitch adjustable electrolarynx", In SP-2004, 761-764, 2004.

[Takahashi, H. et al., 2001] H. Takahashi, M. Nakao, T. Ohkusa, Y. Hatamura, Y. Kikuchi, and K. Kaga, 2001. Pitch control with finger pressure for electrolaryngial or intra-mouth vibrating speech. Jp. J. Logopedics and Phoniatrics, 42(1), 1-8.

[Saikachi, Y. et al., 2009] Y. Saikachi, "Development and Perceptual Evaluation of Amplitude-Based F0 Control in Electrolarynx Speech", Journal of Speech, Language, and Hearing Research Vol.52 1360-1369 October 2009.

[Uemi, N. et al., 1994] N. Uemi, T. Ifukube, M. Takahashi and J. Matsushima, "Design of a new electrolarynx having a pitch control function", In Proceedings of 3rd IEEE International Workshop on Robot and Human Communication, RO-MAN p.198-203, Nagoya, Japan, July 18-20, 1994.

[Nakamura, K. et al., 2010] K. Nakamura, T. Toda, H. Saruwatari and K. Shikano, "The use of air-pressure sensor in electrolaryngeal speech enhancement", INTERSPEECH, p.1628-1631, Makuahari, Japan, Sept 26-30, 2010.

[Fuchs, A. K. et al., 2012] A. K. Fuchs and M. Hagmüller, "Learning an Artificial F0- Contour for ALT Speech", INTERSPEECH, Portland, Oregon, Sept. 9-13, 2012.

[Kubert, H. L. et al., 2009] H. L. Kubert, "Electromyographic control of a hands-free electrolarynx using neck strap muscles", J Commun Disord. 2009 May-Jun;42(3):211-25.

[Poslad, S. et al., 2000] S. Poslad, P. Buckle, R. Hadingham, The FIPA-OS agent platform: Open Source for Open Standards. In Proceedings of Autonomous Agents AGENTS-2000, Barcelona, 2000.

[Argente, E. et al., 2004] E. Argente, A. Giret, S. Valero, V. Julian, V. Botti, Survey of MAS Methods and Platforms focusing on organizational concepts. In: Vitria, J, Radeva, P and Aguilo, I (ed) Recent Advances in Artificial Intelligence Research and Development, Frontiers in Artificial Intelligence and Applications: 2004, pp. 309–316.

[McCabe, K. L. et al., 1995] F.G. McCabe, K. L. Clark. APRIL—Agent Process Interaction Language. In Proceedings of the workshop on agent theories, architectures, and languages on Intelligent agents (ECAI-94), Michael

J. Wooldridge and Nicholas R. Jennings (Eds.). Springer-Verlag New York, Inc., New York, NY, USA, 1995, 324-340.

[Bordini, J. F. et al., 2005] R.H. Bordini, J. F. Hübner, R. Vieira. Jason and the Golden Fleece of agent-oriented programming. In Bordini, R. H., Dastani, M., Dix, J., and El Fallah Seghrouchni, A., eds. Multi-Agent Programming: Languages, Platforms and Applications. Springer-Verlag. Chapter 1, 2005, pp. 3-37.

[CommonWell Project, 2010] CommonWell Project. (2010). [http://commonwell.eu/index.php]. Accessed in February 2014.

[Monami project, 2010] Monami project. (2010). [http://www.monami.info/]. Accessed in February 2014.

[DISCATEL, 2010] DISCATEL. (2010). [http://www.imsersounifor.org/proyectodiscatel/]. Accessed in February 2014.

[INREDIS, 2011] INREDIS. (2011). [http://www.inredis.es/]. Accessed in February 2014.

[INCLUTEC, 2011] INCLUTEC. (2011). [http://www.idi.aetic.es/evia/es/inicio/contenidos/documentacion/documentacion_grupos_de_trabajo/contenido.aspx]. Accessed in February 2014.

[Matsui, K. et al., 2002] K. Matsui, et al., "Enhancement of Esophageal Speech using Formant Synthesis", Journal of Acoustical Society of Japan (E) 23, 2 pp.66-79, 2002.

[Zato, C. et al., 2012a] C. Zato et al., "Platform for building large-scale agent-based systems" 2012 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), pp. 69-73, 17-18 May 2012.

[Zato, C. et al., 2012b] C. Zato, G. Villarrubia, A. Sánchez, I. Barri, E. Rubión, A. Fernández, C. Rebate, J. A. Cabo, T. Álamos, J. Sanz, J. Seco, J., J. Bajo, J. M. Corchado, PANGEA - Platform for Automatic coNstruction of orGanizations of intElligent Agents. In Proceedings of the DCAI, Vol. 151, Springer, 2012, pp. 229-239.

[Zato, C. et al., 2012c] C. Zato, A. Sánchez, G. Villarrubia, J. Bajo and S. Rodríguez "Integration of a proximity detection prototype into a VO developed with PANGEA". 20th International Symposium on Methodologies for Intelligent System (ISMIS 2012), 5th December 2012, Macau (China).

[Matsui, K. et al., 2013] K. Matsui, et al., "A preliminary user interface study of speech enhancement system", Proc. of the 1st International Conference on Industrial Application Engineering 2013, 53-56

[Fujisaki, H. et al., 1988] H. Fujisaki, In Vocal Physiology: Voice Production, Mechanisms and Functions, Raven Press, 1988.

[Sánchez, A. et al., 2013] A. Sánchez et al., "The gateway protocol based on FIPA-ACL for the new agent platform PANGEA" 2013. 11th International Conference on Practical Applications of Agents and Multi-Agent Systems. In Trends in Practical Applications of Agents and Multiagent Systems (pp. 41-51).

[Matsui, K. et al., 2013] K. Matsui, et al, "Development of Electrolarynx with Hands-Free Prosody Control", The Proc. of the 8th ISCA, pp.273-277, Aug.31, 2013.