Research Journal of Applied Sciences, Engineering and Technology 13(8): 611-616, 2016
DOI:10.19026/rjaset.13.3046
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: May 25, 2015 Accepted: July 26, 2015 Published: October 15, 2016

Research Article

Real Time Talking System for Virtual Human based on ProPhone

Itimad Raheem Ali, Ghazali Sulong and Hoshang Kolivand
MaGIC-X (Media and Games Innovation Centre of Excellence), UTM-IRDA Digital Media Centre
Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

Abstract: Lip-syncing is a process of speech assimilation with the lip motions of a virtual character. A virtual talking character is a challenging task because it should provide control on all articulatory movements and must be synchronized with the speech signal. This study presents a virtual talking character system aimed to speeding and easing the visual talking process as compared to the previous techniques using the blend shapes approach. This system constructs the lip-syncing using a set of visemes for reduced phonemes set by a new method named Prophone. This Prophone depend on the probability of appearing the phoneme in the sentence of English Language. The contribution of this study is to develop real-time automatic talking system for English language based on the concatenation of the visemes, followed by presenting the results that was evaluated by the phoneme to viseme table using the Prophone.

Keywords: Phoneme, prophone, real-time talking, virtual character, visemes

INTRODUCTION

The most natural way to conveying the ideas of the personality of a virtual character is through the speech. Therefore, the researchers focus on the visual speech methods to represent the interaction between the virtual character and human's behavior. However the speech is not composed of sounds, but also many emotions, facial expressions and the corresponding articulatory movements. All of these corresponding provide an important impact on the credibility of the virtual character. If the virtual talking is not conducted smoothly or the lip synchronization does not match the sound that means the human will find awkward animation.

Talking is commonly represented as a sequence of phonemes. Each phoneme can be associated with a viseme (i.e., the visible shape of the vocal articulators with the lips, teeth, tongue, jaw and cheeks). All talkers have influenced the production of a given phoneme but not all visible, therefore different phonemes may be associated with the same viseme. In facial animation, the talking animating events previously created each viseme. Later, they concatenate the visemes according to the talking they want to animate using interpolation schemes. So, manual speech animation is tedious and a waste of time. Therefore, a lot of efforts in automatic approaches have been employed for synchronizing the talking (Schuller *et al*., 2013).

The visual talking can be divided to cases according to the way of the input speech and the articulatory movements. In the first case of phoneme to viseme mapping, the input text, or audio is analyzed to obtain the phoneme and its alignment then mapped to visemes to be organized with time aligned. The time between the phoneme and visemes is important for the end result because if there's anything lacking, the animation will have an unexpected visual effect appearance. Actually, a good timetable is not sufficient for a real-time talking animation. Thus, the diphones and triphones techniques have been proposed to solve this problem using large visual talking databases (Zhou *et al*., 2010). The main advantage from the automatic systems is making the visual talk faster and there is more accuracy that comes from not needing the individual phonemes to be identified as the continuous signal approaches. Actually, the problem of the sub-phonetic case is highly sensitive to the noise. However some works of this area have been done in other languages, such as (Akagunduz *et al*., 2004; Balci, 2004; Esposito and Esposito, 2011; López-Colino and Colás, 2012; Gonseth *et al*., 2013).

The challenging task in virtual talking animation is the selection of the visemes synchronized with the audio and the modelling of articulation. López-Colino and Colás (2012) and Serra *et al*. (2012) introduced a system to create a modular virtual talking animation framework. The contribution of this study involves a new concept in virtual talking animation called ProPhon. Prophon is a canonical set of visemes for the pairwise phoneme reduced by the original phoneme set, used to minimize the number of pose of 3D face model.

Corresponding Author: Itimad Raheem Ali, MaGIC-X (Media and Games Innovation Centre of Excellence), UTM-IRDA Digital Media Centre Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

Table 1: Phoneme-to-viseme mappings

Viseme class	Phonemes
A	/m/, /b/, /p/
B	/f/, /v/
C	/d/, /n/, /t/
D	/s/, /z/
E	/ʃ/, /ʒ/
F	/ɾ/
G	/l/, /ʎ/, /ɲ/
H	/g/, /k/
I	/R/
J	/ĩ/, /j/, /i/
K	/ɨ/
L	/ɐ/, /ã/, /e
M	/, /ɛ̃/
N	/a/, /ɛ/
S	/ɔ/, /o/, /õ/, /u/, /ũ/, /w/ silence/neutral

Table 2: Set of English and selected phonemes (TRueSpel, 2001; Scobbie *et al*., 2006)

English phoneme (40)	Phoneme set (15)	Example	Percentage (%)
ae, ah, ax,	ah	Cat, cut, ago	2.1
Ao, Ow, aw, O	oh	Dog, go, Foul	1.2
Ey, ee, Ay	Ee	Ate, pet, bite	5.0
Ih, iy	I	Feel, fill, debit	9.9
W, uw, uh	W	With, too, book	1.2
Aa	Aa	Father	1.9
H	H	Help	0.7
R, er	R	Red, fur	3.9
Sh, ch, jh, zh	Sh	She, chin, joy, pleasure	1.2
Th, dh	Th	Thin, then	0.3
F, v	F	Fork, vat	1.7
D, t, s, z, l	D	Dig, talk, sit, zap, lid	4.0
K, g	K	Cut, gut	2.6
P, m, b	P, m	Put, mat, big	3.3
N, ng	N	No, sing	7.8

Phoneme to viseme mapping: Phoneme is the smallest unit of the speech that can contrast the utterances. In a phonetic speech, combining phonemes, rather than the actual letters, creates words. For example, in the word "foot", the "oo" sound would be represented by the "UH" phoneme. The phonetic spelling of the word would be "F-UH-T". When the phonemes are spoken, the mouth changes shape to form the phonemes that are heard, not specifically the words that are spoken. The words "bye" and "pie" have different meaning and are pronounced as (bai, pai) respectively. They are two distinct phonemes in these words. Visemes is the visual phoneme that describes the facial and the pose during the articulation. Visemes in facial animation has a relation on "one to many" with the phonemes. This section describes the phoneme to the Viseme mapping of the 35 phoneme set to 15 classes using Prophone method, that reduced the visemes of the phonemes depending on the probability of appearing the phoneme in English language, as shown in Table 1, depending on the place and manner or articulation and Table 2 show the Prophone. These classes are divided depending on the classification of the vowels (back, close front, close central, close mid front open mid...) with present to the neutral stance, or silence. The visemes created by a digital artist, who has a good experience based on the articulatory movements when pronouncing the phoneme alone and pronouncing in the context of the sentence.

Real-time talking system description: A virtual Talking character to synchronize the facial movements with an audio of utterance becomes impractical if the length of this utterance increases. The proposed system handles the utterances of different lengths automatically based on a new method for visual talking animation. Figure 1 show the data process. The input to the system is a text/audio file. The proposed system generates a sequence of animations that are associated with the developed visemes set of phoneme. Therefore, instead of having animators create animations from scratch the Prophone is represented as a set of animation using a canonical set of face poses. This not only simplifies the task for animators but also allows being portable to different animated characters in the future. Prophone can be applied on that character to generate high quality lip syncing animations. The method constructs facial movements by combining several static face poses over time to produce set of Prophone animations.

Visemes are the visual counterpart of speech animation generation. The morph targets are modelled for a set of phonemes in such a way that the agent acquires proper lip synchronization with the audio. Generally, there are 15 phonemes used by TTS for generating the audio, which are selected from the 40 phoneme in English language. Therefore, the subsequent morph targets for our 3D model are created for all the phonemes as summarized in Table 2 Forty English phonemes mapped to our common set of

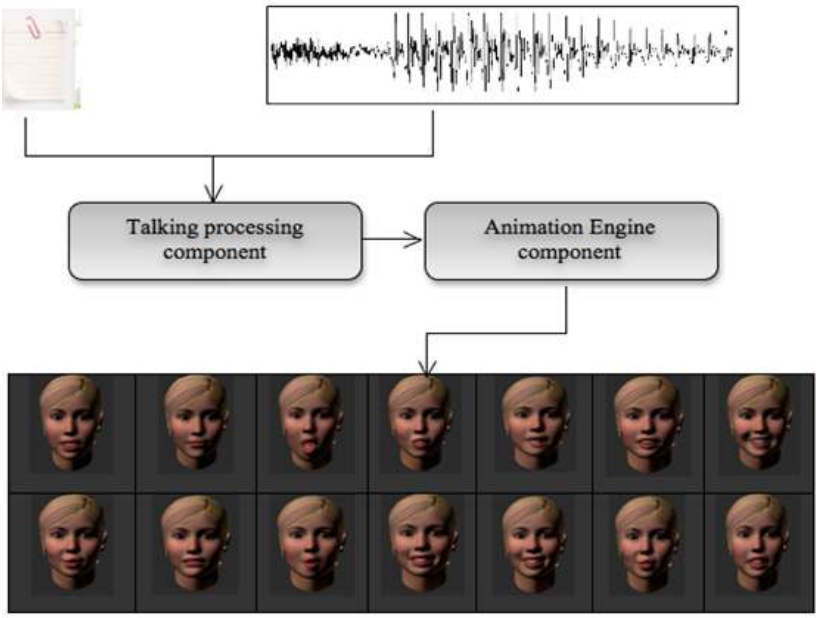

Fig. 1: The overall of the system and the facial poses for the visual phoneme

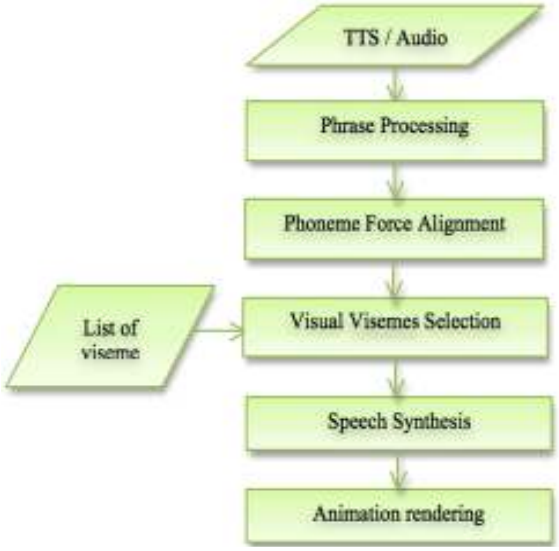

Fig. 2: The system architecture flowchart

phonemes, the percentage of the phonemes gets from these three sites (TRueSpel, 2001; Scobbie *et al*., 2006). The left column lists the full set of English phonemes. The 2nd column lists all the pairwise phoneme combinations for the given phoneme schedules. The 3rd column lists all the canonical visemes. The 4th column lists common set phoneme pair and 5th column list the frequency of common phoneme set generated from TTS engine on a corpus of approximately 200 utterances. Accurate selection of phoneme and visems together with their mapping and alignment play paramount role in the animation. A brief description on these issues is emphasized below.

As aforementioned, the goal of this system is to construct a set of animations that are associated with the co-articulation. In this regard, a new technique called "ProPhone" is suggested for choosing phoneme pairs as a canonical unit of animation because they allow the effects of co-articulation effects that are not possible by associating animation with individual phonemes. In articulation science, there are two terms named Diphones and Triphones. Diphones are commonly used during machine learning techniques, which represent timings between the middle of one phoneme and the next. Moreover, an animation of a Prophone is intuitive whereas diphone is non-intuitive representation.

The visual talking animation processes begin with the neutral pose with the 3D visemes and the utterances that are given to be talked. The utterance is the direct input to the visual talking system. Figure 2 illustrates the conceptual architecture of the proposed system. A data-driven approach to visual talking synthesis requires a database of audio-visual recordings of a talking face that captures the different phonetic combinations in the language being used. The data needs to be processed both in the audio and visual domains in order to obtain a representation that is suitable for use with Machine Learning algorithms. We begin by the audio aligned with speech using Microsoft SAPI (NET Framework Conceptual Overview, 2012). The data corpora used in this system are then presented, followed by details of visual and audio processing.

The proposed system consists from two main components: first one the interface to the virtual character animation engine with the user. And the second one the processing of talking, that analyses the

utterance and generate the data that drives the animation. The proposed system is a clear distinction between the talk processes of the input utterance and the animation of the virtual character. Through the adaptation of our system, it has integrated different animation engines.

The interface animation engine component: The animation engine explores the input data then it is translated into the final animation by the rendering engine. The proposed system, which uses Xface, 3D modeling, is an open source project (Balci, 2004). A virtual character, which was created in Xface, changed the neutral pose to create the visemes that are then combined based on the utterance given by the talking process component.

Talking processing component: This component deals with the creation of the animation. The input data (text, audio, or both) to the system will drive the virtual character animation. The processing of the Phrase processing obtains the phonetic linguistic of the input utterance. Our system uses Microsoft Speech API (SAPI 5.4) Microsoft SAPI (NET Framework Conceptual Overview, 2012) together with the phonetic lexicon developed at Microsoft. The Phoneme Force Alignment module handled the synchronization of the visemes and the audio data with guarantees so that talking is matched correctly with the lips motions. The visualization of the character is very important because the animated utterances become incomprehensible, if it not correct. Our system used Automatic Speech Recognition (ASR) and text to speech synthesis (TTS) that are commonly used for synchronization. The speech synthesis module is used when the input data is text. With ASR, we can obtain the time aligned phonetic linguistic. ASR aligns a speech signal with the phonemes in the utterance. The time duration of the phoneme in the utterance should be the same as the duration of the audio signal. The visual selection module is responsible for choosing the viseme blend shape by the phoneme-to-viseme mapping. Our system maps the phoneme to the corresponding visemes. The viseme were created directly observing the mouth movements of a person speaking each phoneme independently.

Implementation and evaluation: The proposed system is evaluated by a preliminary subjective user evaluation to analyze the effects of the new mapping of phoneme-to-viseme. The next section illustrates the experimental evaluation and its result.

Experiment: The evaluation was executed using 40 persons (between 20 and 65 years old) at a multimedia systems lab at Universiti Technologi Malaysia (UTM). All of them did not have any problem in the vision or hearing, but three of them had expertise in this area. The evaluation of the proposed system was carried out using three sentences presented:

S1 : Communication is important to convey the ideas
S2 : Messi plays football with Renaldo
S3 : The man does not see ships in the sea

Each sentence was animated using the phoneme-to-viseme mapping in below section. The animations of the virtual character were shown twice to the persons who then filled out a questionnaire. Conforming to the mapping, the persons had to select one of the following alternatives for each sentence:

A1 : Strongly the animation is clear
A2 : A bit the animation is clear
A3 : Neutral

In the present work, a facial model is designed especially for animating talk. The animation of talking requires highly deformable lips, tongue, chin because mouth conveys important information about the meaningful communication signal. A facial model is desired to represent the precise geometric aspects of any human face.

The experimental analysis and results are divided in two sections. The first one is dedicated to the quantitative analysis that is carried with the proposed system associated with some of the most popular Virtual Human Character such as Xface (Balci, 2004).

RESULTS

The articulation of a phoneme determines the degree of emphasis received by the visual phoneme together with the duration. The first stage in the present system is the input of neutral audio or text that created speech. This section explains the examples of audio modifications with stuttered speech. Figure 3 depicts the waveform, corresponding phonemes and words for the sentence “when you are happy...” that is spoken neutrally using Praat software (Boersma and Weenink, 2001). Figure 4 presents the signal of audio waveform and the corresponding speech segments of the same sentence. Figure 5 illustrates the English visual phoneme articulation with normal expression for the sentence “the natural beauty”, with frontal side.

The comparison of the preferences for the sentences during the experiment is shown in Fig. 6.

In Fig. 3, we can see the persons slightly preferred the decreasing in mapping for S1 and S3 but strongly to S2. Our system particularly is favored more than Xface under this mapping. We can deduce that the differences of visemes classes had little importance for the quality of S1 and S3, but had a positive effect on the quality of S2. That means some sentences have a small animation effect without changing on the quality. Actually, most

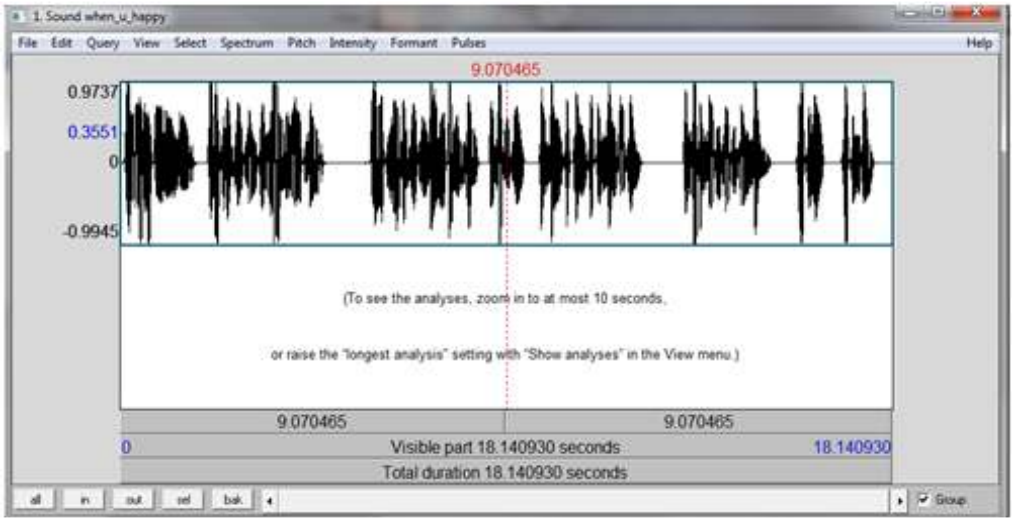

Fig. 3: The speech signal of the sentence "when you happy in your inside, the natural beauty will shine in your inside"

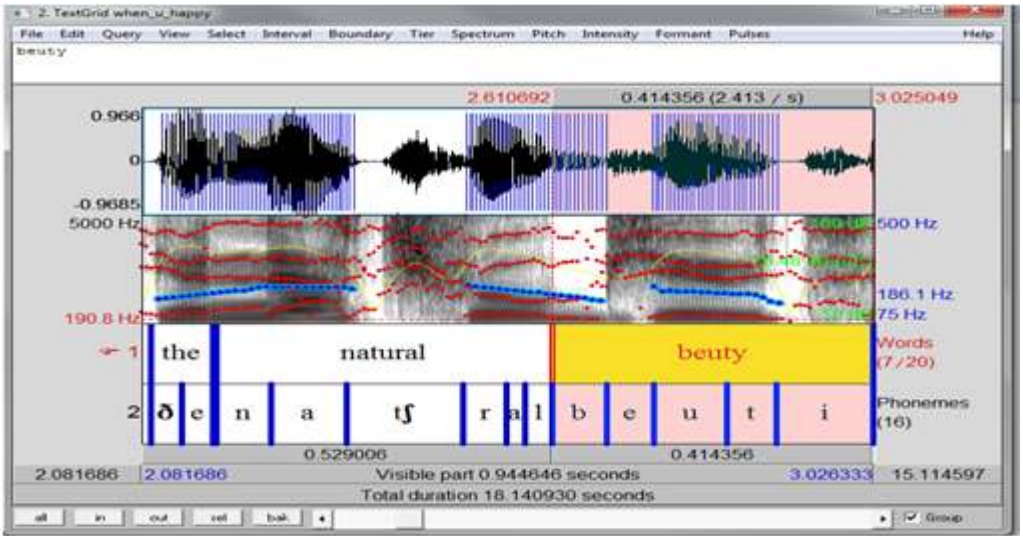

Fig. 4: Screen capture from Praat software showing the speech signal and the phoneme-word audio segmentation of the neutrally spoken sub-sentence "the natural beauty"

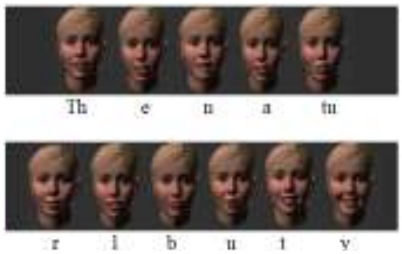

Fig. 5: English visual phoneme articulation with normal expression for the sentence "the natural beauty", in frontal side

of the preferences are focused on the more neutral alternatives A2 and A3 greatest in case of S1 (1.3), S2 (2.40) and S3 (2), respectively. This preference shows that the different mapping does influence on the quality

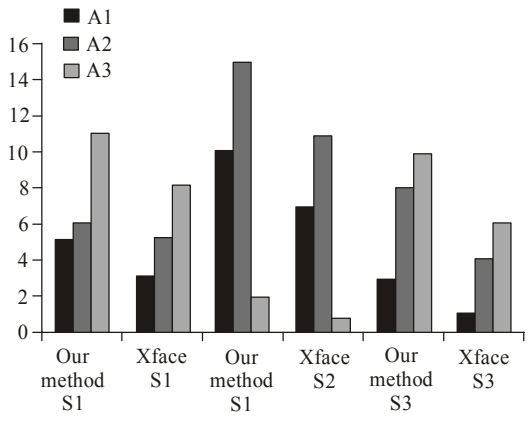

Fig. 6: The comparison preferences for three sentences between our method and Xface

of talking animation as expected and is used by another researcher (Serra *et al*., 2012; Wang *et al*., 2012; Zhang *et al*., 2013; Xu *et al*., 2013).

CONCLUSION AND RECOMMENDATIONS

Visual talking system is a challenging system that creates 3D character based on text or audio or both input to obtain a virtual character like a human's facial movements. This study presents a modular of a virtual talking character system aimed to speeding and easing the visual talking process as compared to the previous techniques. However, speech synthesis systems are focused on translating written text into audible speech with the primary goal to achieve speech-to-speech technology. The voice recognition text-to-audio with speech and translation technologies is integrated to create a pragmatic virtual human character. It is asserted that with speech-to-speech in real time, two human talking in languages can understand each other. A realistic talking system is established by synchronize the lips motions with the phonemes.

The contribution of this study is to create a real-time system has the capability of generating a visual talking of the English language with more realism. The evaluation experiments show that the quality of the animation is still an inconclusive change, according on the articulation the sentence. In the near future we intend to improve the reality of this system and finding a solution to the articulation problem. We strongly intend to improve this system by adding the emotional facial expression to it, which gives makes it more real just like a human's facial expressions.

ACKNOWLEDGMENT

This research is supported by the Ministry of Science and Technology (MOSTI) collaboration with Research Management Center (RMC) and UTM-IRDA Digital Media Centre of Excellence MaGIC-X (Media and Game Innovation Centre of Excellence), Universiti Teknologi Malaysia (UTM).

REFERENCES

Akagunduz, E., U. Halici and K. Ulusoy, 2004. Simulation of Turkish lip motion and facial expressions in a 3D environment and synchronization with a Turkish speech engine. Proceeding of the IEEE 12th Signal Processing and Communications Applications Conference, pp: 276-279.

Balci, K., 2004. Xface: MPEG-4 based open source toolkit for 3D facial animation. Proceeding of the Working Conference on Advance Visual Interfaces, pp: 399-402.

Boersma, P., and D. Weenink, 2001. Praat 3.9. 15 [Computer Software]. Institute of Phonetic Sciences, Amsterdam, the Netherlands.

Esposito, A. and A.M. Esposito, 2011. On Speech and Gestures Synchrony. In: Esposito A. *et al*. (Eds.), Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 6800: 252-272.

Gonseth, C., A. Vilain and C. Vilain, 2013. An experimental study of speech/gesture interactions and distance encoding. Speech Commun., 55(4): 553-571.

López-Colino, F. and J. Colás, 2012. Spanish sign language synthesis system. J. Visual Lang. Comput., 23(3): 121-136.

NET Framework Conceptual Overview, 2012. Microsoft Developer Network Platform, Retrieved from http://msdn.microsoft.com/enus/library/w0x726c2%28v=vs.90%29.aspx.

Schuller, B., S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller and S. Narayanan, 2013. Paralinguistics in speech and language—State-of-the-art and the challenge. Comput. Speech Lang., 27(1): 4-39.

Scobbie, J.M., O.B. Gordeeva and B. Matthews, 2006. Acquisition of Scottish english phonolgy: An overview. Proceeding of QMUC Speech Science Research Centre Working Paper WP-7, Queen Margaret University College, 7: 3-30.

Serra, J., M. Ribeiro, J. Freitas, V. Orvalho and M.S. Dias, 2012. A Proposal for a Visual Speech Animation System for European Portuguese. In: Toledano, D.T. *et al*. (Eds.), Advances in Speech and Language Technologies for Iberian Languages. Springer-Verlag, Berlin, Heidelberg, 328: 267-276.

TRueSpel, 2001. English-Truespel (USA Accent) Text Conversion Tool. Retrieved from: http://www.foreignword.com/dictionary/truespel/transpel.htm.

Wang, L., H. Chen, S. Li and H.M. Meng, 2012. Phoneme-level articulatory animation in pronunciation training. Speech Commun., 54(7): 845-856.

Xu, Y., A.W. Feng, S. Marsella and A. Shapiro, 2013. A practical and configurable lip sync method for games. Proceeding of the Motion on Games (MIG'13), pp: 131-140.

Zhang, L., M. Jiang, D. Farid and M.A. Hossain, 2013. Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot. Expert Syst. Appl., 40(13): 5160-5168.

Zhou, Z., G. Zhao and M. Pietikäinen, 2010. Synthesizing a talking mouth. Proceeding of the 7th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'10), pp: 211-218.