A Review of Speech Processing and Facial Animation Technologies for Solving Thai Word Mispronunciations

Panyayot Chaikan
Department of Computer Engineering, Faculty of Engineering
Prince of Songkla University
Hadyai, Songkhla, Thailand
panyayot@coe.psu.ac.th

Abstract— **This paper presents a comprehensive review of Thai speech processing and facial animation technologies for solving the problem of mispronunciation. The focus is on the detection of pronunciation errors in Thai words containing the initial consonants /r/, /l/ and the consonant clusters /pr/, /phr/, /tr/, /kr/, /khr/, /pl/, /phl/, /thr/, /kl/, /khl/, /kw/, and /khw/. Open issues and research directions are also suggested.**

Keywords- Pronunciation scoring, Pronunciation error detection, Speech recognition, Thai.

I. INTRODUCTION

Mispronunciation is one of the most common problems of Thai language teaching, especially for people who do not use Thai as their first language, e.g. hill tribes, people living in Thailand's three southernmost provinces [9], [21], [10], and foreigners [11], [12]. In fact, good pronunciation is a problem for Thais of all ages and genders [5]. The main reason is that some syllables are difficult to pronounce, and can easily be mispronounced as similar, simpler syllables. For example, the word "รัก" should be pronounced as /rak/, but is often spoken as /lak/. This type of mispronunciation is common because most people think it is not a serious problem.

Mispronunciation also leads to incorrect spelling, for example, students who mispronounce the word "กราบ": /kraab/ as /klaab/ or /kabb/ frequently write it as "กลาบ" or "กาบ". This problem often occurs amongst students below the third grade of primary school in Thailand [8], [3].

Error of pronunciation can be divided into four types: substitution, omission, distortion, and addition [20]. Substitution errors are the most common in Thai language pronunciation, as shown in Table 1.

Many attempts have been made to solve the problem of Thai language mispronunciation [1]-[9], relying on the use of lesson plans in conjunction with a variety of materials, such as cards, songs, rhymes, poetry, and games. Initially, they use basic words, and the complexity of the words and level of difficulty are gradually increased. The result is a significant improvement in the pronunciation quality of the students, but the techniques require a high level of interaction between the teachers and the students. Reinforcement by parents is helpful but only occurs if they have a good grasp of Thai pronunciation. For this reason, more recent research has used electronics media, such as CDs and cassette tapes, to give students the opportunity to listen and learn by themselves at home [3]. A computer can be utilized to overcome the interactive problem inherent in CDs and tapes, since software can allow a student to re-listen to words as many times as they want [7], [8]. However, typical software plays pre-recorded voice files, and lacks human voice analysis for scoring the student's voice. This makes it impossible for a student to assess and improve their pronunciation with software alone. However, if speech processing technologies are utilized, then automatic pronunciation scoring and error detection does become possible.

The rest of this paper is organized as follows: section 2 overviews Thai speech recognition applications, section 3 outlines the future requirements for pronunciation training software, section 4 describes a feature extraction method for speech data and its recognition mechanisms, and section 5 introduces acoustic modes developed for Thai. A review of the speech corpus developed by Thailand's National Electronics and Computer Technology Center (NECTEC) is provided in section 6. Section 7 examines pronunciation scoring and error detection, and facial animation is explained in section 8. Conclusions and discussion are provided in section 9.

II. SPEECH PROCESSING IN THAILAND

Speech processing systems can be divided into two broad groups: if separated words are required during the speech processing setup then the system is utilizing isolated word speech recognition, otherwise it is considered to be a continuous speech processing system.

In terms of applications, Thai speech processing research can be divided into four categories:

1) *Speaker Verification:* speech is employed as an authorization token to decide whether a person can be allowed to use a computer-based system or not [61]. One-to-one comparison is made between the enrolled features and the input voice, and the system either rejects or accepts that the input and enrolled voices come from the same person.

This work is partially supported by the Faculty of Engineering, Prince of Songkla University, under grant no. ENG-53-2-7-2-0107-S.

2) *Speaker Identification:* speech is used to identify the person who wants to gain authorized entry into a system [14, 16]. This differs from speaker verification is that the identification system supports multiple users. A one-to-many comparison is made between the input voice and enrolled templates stored in a database. The best matching voice candidate is obtained, and the input voice is accepted only if the matching score is above a specified threshold.

3) *Speech-to-Text:* speech is automatically translated into text, as in a voice command system for controlling electrical appliances [13], [15]. This speaker-independent system is not restricted to a specific individual, but is capable of translating anyone's voice. For high recognition accuracy, large amounts of speech data from people of different ages and sexes is required for its training.

4) *Text-to-Speech:* a text-to-voice system analyses input text and assigns phonetic transcriptions to words, to synthesize a concatenation of phonemes. The VAJA [28] text-to-speech engine developed by NECTEC has an open API. jRAJA [65] is an improved version of VAJA which is capable of synthesizing Thai text even if some input words are not in the system dictionary. Commercial Thai text-to-speech software called Salika [64] is also available.

Fundamental research on Thai mispronunciation detection and scoring is required before software can be developed aimed at pronunciation practice.

TABLE I. FREQUENTLY INCORRECT PRONUNCIATIONS IN THAI

Type	Description
I	Replacement of /r/ with /l/ and vice versa
II	Replacement of /kr/, /khr/, /pr/, and /phr/ with /kl/, /khl/, /pl/, and /phl/ respectively
III	Replacement of /kr/, /khr/, /tr/, /pr/, /phr/,/kl/, /khl/, /pl/, and /phl/ with /k/, /kh/, /t/, /p/, /ph/, /k/, /kh/, /pl/, and /ph/ respectively
IV	Replacement of /khw/ with /f/

III. PRONUNCIATION TRAINING SOFTWARE

Software for Thai pronunciation practice should include the following features:

1) *Automatic Scoring:* when a students does an exercise or test, the word scores must be collected and reported. The next chapter or section becomes available only if the previous chapter or section has a satisfactory score.

2) *Error detection and suggestion:* the ability to detect errors, along with a score, gives a student the chance to solve their pronunciation problems. For example, when a student mispronounces word /รัก/ as /lak/, the system may show the recognized phoneme and a correct pronunciation can be suggested to the student. This helps a student to pinpoint his/her weaknesses of pronunciation.

3*) Noise immunity:* most speech data used for testing and training is recorded in a controlled environment with low levels of noise [25]. However, in a real user environment, noise is unavoidable, and [15] states that their system cannot function correctly if the noise level is more than 30 dB. Only two Thai publications discuss noise cancellation [39], [63]. The noise can be suppressed if a good filter is applied to the noisy input signal, and for this reason, automatic background noise detection must be further studied.

4) *Correct pronunciation demonstration*: the ability to play each word through a loudspeaker is indispensable, because students can re-listen to word as many times as they want. There are two ways of generating the sound: the first relies on playing sound files of the respective words. This approach is inconvenient to update because sound files must be recorded when new words are added. The second approach, text-to-speech, answers this problem by automatically synthesizing the input words as a voice. For this reason, VAJA [28] or jRAJA [65] are recommended for this purpose.

5) *Realtime Response:* the instantaneous response to a student's voice means that they can receive feedback on pronunciation immediately. The processing speed of the scoring algorithm is a concern, especially in systems with low-to-medium computational capabilities such as mid-level mobile phones, tablets, or netbooks. However, on PCs, high-end mobile devices, or notebooks, this is not a critical factor due to their superior performance.

6) *Correct lip-movement demonstration:* a demonstration helps the student learn how to move their lips and tongue to obtain the correct pronunciation. This feature also helps the listening impaired to pronounce words more effectively, and recorded lip movement videos encourages students to practice by themselves. However, when the training vocabulary changes, it is inconvenient for the teacher to record new lip movement videos. Text-to-visual-speech systems, such as [60], [65] are an interesting way to solve this problem.

7) *Attractive to children:* the user interface should be user-friendly, good-looking, and easy to use. Software-based exercises or test may become tedious and boring, especially for primary school students; the addition of relevant games is one possible answer.

8) *Upgradeability:* the database of training exercises, games, and test vocabularies should be easily updated by teachers or parents

IV. FEATURE EXTRACTION OF SPEECH AND AUTOMATIC RECOGNITION ALGORITHMS

Human speech research typically employs 16-bit quantization level data sampling, at two different rates. For example, [14], [18], and [19] uses 11.025 KHz while [17], [25], [27], [29] employs 16 KHz. Frequently, LPC or MFCC feature extraction is applied to the digitized signal.

LPC (linear predictive coefficient) utilizes a mathematical model to approximate the human vocal tract [18] as tube diameter variations, which allows precise feature extraction of speech data.

MFCC (Mel-frequency cepstral coefficient) uses a spectral envelope which differs from conventional cepstrum in terms of the space between its frequency bands. It is nonlinear, focusing on lower frequencies over higher ones, in a similar way to human perception.

LPC is utilized by [16], while MFCC is employed by [41-48]; other features, such as PFL [14] are rarely seen. MFCC is more commonly used than LPC due to its superior recognition accuracy [17], and MFCC also gives higher accuracy than PFL [14]. Variant types of MFCC are employed for Thai, including MFCC+D, MFCC+DA, and a modified MFCC proposed in [26]. MFCC+DA seems to be the most popular in research at present.

One trend in pattern recognition applications is their multimodal basis where many types of features are combined to increase the recognition performance. However, almost no published research employs multi-type features, except for [62] which combines MFCC and LPC to increase the recognition accuracy.

As described above, if only one feature is utilized, then MFCC is the most suitable for pronunciation error detection and scoring. Unfortunately, the use of multiple features to improve accuracy has not attracted much attention for Thai.

Three popular recognition mechanisms for Thai speech processing are: Hidden Markov models (HMM), artificial neural networks (ANN), and dynamic time warpings (DTW). HMM is utilized by [17, 20, 23, 24, 40-49], DTW in [14-16], [56], and neural networks in [51-55].

HMM is suitable for problems that require a sequence of decisions, where each decision is influenced by a previous decision, which is a typical scenario for speech data. A model consists of hidden states, transitional probabilities between those states, with emission probabilities for each state obtained through training. HMM for speech processing is based on a left-to-right model [18].

DTW seeks an optimal alignment between two given sequences, by nonlinearly warping one sequence to match the other. Since similar speech may vary in speed, DTW is a suitable approach for measuring the similarity between two speech sequences. DTW is mostly used for isolated word recognition, but it can be employed for continuous speech with some modifications.

An ANN model consists of nodes called neurons inspired by the functionality of biological networks. During supervised learning, its biases and weights are set through training. Speech processing schemes using ANN may be static or dynamic. A static ANN is applied to all of the input speech at once, and a decision is made. In a dynamic ANN, a small processing window slides over the input speech, and a decision is made based on the changing window.

The accuracy of isolated numeral recognition systems using HMM and neural networks was studied in [19, 56], and HMM supersedes other techniques in terms of accuracy. Although, DTW gives better accuracy than ANN, it is much slower [14].

V. Acoustic Model for Thai Language

Automatic speech recognition can be implemented in terms of syllable or phoneme units. Syllable units are easy to implement, but are inappropriate for large vocabularies [51]. As a consequence, phoneme units have been adopted for Thai speech recognition.

There are two general forms of Thai syllables: /C_i+V/ and /C_i+V+C_f/. The C_i stands for the initial consonant, V represents a vowel, and C_f is the final consonant. The acoustic model proposed by [22] comprises of 21 single initial consonants, 12 double initial consonants, 24 vowels, and 12 final consonants, totaling 69 models altogether. Thai phonetic units made up of 35 phonemes is proposed by [49]. A reduced acoustic model consisting of 41 phonemes appears in [17], made possible by combining similar phonemes to reduce recognition errors.

Thai is a tonal language utilizing 5 tones: mid, low, fall, high, and rise [17], [22]. Phonemes with tone modeling were proposed by [22], and they reported that the model without tone modeling yields higher accuracy than the model with tone modeling. This result was confirmed by [17].

Pronunciation is graded by syllable segmentation with each matching segmented phoneme being assigned a score. Most Thai speech processing papers reported their results in terms of overall recognition accuracy, and none utilize segmentation accuracy.

In a pronunciation training system that focuses on /r/, /l/, and their cluster consonants, tone modeling is not as important as the accuracy of phoneme segmentation.

VI. Thai Speech Corpora

Prior to 2002, Thai speech processing researchers used their own speech corpora for training and testing their systems, whch meant that the performance of one system could not be compared to others [28]. In 2003, NECTEC introduced the first standard speech corpus for performance evaluation, and there are now six Thai speech corpora: NECTEC-ATR, LOTUS [31], LOTUS-Cell [32], VoiceCom, NECTEC-TRUE, and LOTUS-BN [29]. Each corpus focuses on a different speech application domain: NECTEC-ATR contains 5,000 isolated words corpus, LOTUS consists of continuous speech, VoiceCom collects voice commands, LOTUS-Cell and NECTEC-TRUE are suitable for testing telephone conversation voices, and LOTUS-BN utilizes broadcast news [22]. However, 67.5 percent of the speech in these corpora were collected from speakers aged 21-25 years [25], [32]. This means that they are not suitable for training and evaluating the performance of pronunciation errors in applications aimed at primary school students, the most suitable age for practicing pronunciation [6]. Speech corpus for pronunciation should be constructed for speakers of appropriate ages, but this has not yet occurred.

VII. Pronunciation Scoring and Error Detection

There is no work on error detection or scoring of Thai language pronunciation, so this section gives a review of techniques developed for other languages.

Phonetic segmentation is required for pronunciation scoring [35]. Segmentations are acquired by means of HMM, and then force-alignment [38] is utilized to obtain the spectral match and duration score. Posterior probability is utilized with phoneme dependent thresholds in [57]. Two scoring algorithms are proposed by [37], and they report that log-posterior probability scores outperform log-likelihood scores. The log-likelihood of input speech is obtained from a Viterbi decoder [58], and the posterior probability is calculated from the log-likelihood score and the language model. The use of more than one recognizer for pronunciation scoring is proposed by [59], where HMM and ANN are utilized for Mandarin.

Lip reading and speech recognition is used to score a learner's pronunciation in [36], along with a video of the teacher's lip movement. This means that the learners can practice alone.

VIII. Facial Animation

A visual text-to-speech engine automatically synthesizes lip-animation videos from input text, which means that pronunciation training system does not need to re-capture new lip-movement videos when new vocabulary is inserted. This technology ranges from simple systems that only create local lip-areas, to sophisticated 3D talking heads. Unfortunately, no research paper has reported on its use for solving Thai language pronunciation.

Text-to-visual-speech is similar to a conventional text-to-speech in that it uses input text to create a series of phonemes and timing information. However, instead of synthesizing the obtained parameter to sound like a conventional text-to-speech engine, sequences of key-frames [67] called visemes are produced. A viseme is an image that represents oral position and shape when each phoneme is pronounced. The challenge of creating realistic and plausible text-to-visual-speech system comes from the fact that the correspondence of each phoneme and viseme is not a simple one-to-one mapping. Moreover, the influence of the previous viseme on the current and next viseme may be required for creating smooth lip-movement transitions. A coarticulation model [68], developed by Cohen and Massaro, is widely utilized to deal with these problems.

There are two ways to create lip movement animations: sample-based and model-based approaches. In sample-based, snippets of lip area images are extracted from the video frames of a person talking, and each image is uniquely labeled and stored in a database. A lip movement animation is obtained by selecting images from the database according to the phoneme derived from the text-to-visual-speech engine. A Viterbi-search algorithm [69] creates a coherent sequence of mouth movement but requires a large number of images in the database. Interpolation and morphing techniques can deal with this problem.

A model-based lip-movement animation usually comes with facial animation. Polygon meshes specify surface shapes, along with physics-based muscle modeling [70], to achieve a sophisticated 3D face model. This approach provides more realistic facial animation than the sample-based technique, but consumes more computation power. In a text-to-visual-speech system, each phoneme is translated into a facial animation parameter according to the face model. More information may be added to the model, such as the movements of the eyes, the eyelids and eyebrows for displaying emotions.

Instead of building a talking head from scratch, the Xface [71] and iFACE [50] toolkits can be used to reduce the development time of a model-based lip movement animation for Thai.

IX. Conclusion and Discussion

Thai speech processing has a long history, but has never been used for pronunciation scoring and error detection. Several technologies can aid in the development of Thai pronunciation training software, including feature selection and extraction mechanisms, text-to-speech engines, and acoustic models. However, basic research on pronunciation error detection, scoring, and noise reduction are still required, and the pronunciation corpus needs further development. The use of text-to-visual-speech translation is an interesting way for students to improve the quality of their Thai language pronunciation.

Acknowledgments

The author is grateful to Dr. Andrew Davison for his kind help in polishing the language of this paper.

References

[1] P. Kgamjit, "A construction of pattern drills for Thai alphabet in reading ร (R), ล(L) and ว(W) cluster words for prathomsuksa 3 students in Udon Thani province," Master Thesis, Burapha University, October, 1990.

[2] S. Tanpiti, "A comparative study of Prathom Suksa II student ability in reading aloud: an emphasis on the clusters with /r/, /l/, and /w/ by using skills-practice games and skills-practice in the teacher's manual," Master project. Srinakarinwirot University, January, 1993.

[3] W. Punyaruang, "Use of songs and games to practice reading words containing consonant clusters with ร[r], ล[l], ว[w] for Prathom Suksa 2 students," Master thesis, Chiangmai University, October 1995.

[4] J. Maksavad, "The effect of modeling technique on the pronunciation of consonant cluster ร ล ว of prathom Suksa II students of Surhaobangmakhue school in Khet Wattana, Bangkok," Master thesis, Srinakarinwirot University, October 2004.

[5] NA, "วิกฤตวิบัติภาษาไทยภาคการออกเสียง [The crisis of Thai language pronunciation]," [online] available: http://www.9digits.com/index.php?option=com_content&viewarticle&id=149&catid=75 blog&Itemid=36.

[6] J. Thongprasert, "ภาษาไทยของเราดี เป็นศักดิ์เป็นศรีของชาติ [Thai language is good and a national prestigious]," ISBN: 974-575-693-8, 2nd edition, Mahachulalongkornrajavidyalaya Press. 2001.

[7] W. Wangwech, "สื่อสารสนเทศแม่แบบช่วยการอ่านออกเสียง ร ล และคำควบกล้ำชั้นมัธยมศึกษาปีที่ 2 [The prototype of E-learning for /r/ and /l/ pronunciation, designed for eighth grade students]," [online] available: http://www.nawama.ac.th/wantanee/

[8] N. Suthiart, "A development of the multimedia computer program on pronunciation of clusters for Pratom Suksa 3 students," Master Thesis, Rajabhat Sakon Nakhon University, March 2006.

[9] D. Chitaree, "Using songs to develop pronunciation of words containing consonant clusters with ร(r), ล(l), ว(w) of Prathom Suksa 4 hill tribe students," Master Thesis, Chiangmai University, April 2003.

[10] K. Kanongdech, "Effects of modeling practices on pronunciation of the bilingual kindergarten students in the three southern border provinces," Master Thesis, Prince of Songkla University, 1995.

[11] S. Narasuwan, "Japanese students' pronunciation problems with initial single-consonant sounds in the Thai language," Master Thesis, Thammasat University, May 2004.

[12] N. M. Phuong, "Problems in pronouncing thai consonants and vowels of Vietnamese students," Master Thesis, Srinakharinwirot University, May 2009.

[13] R. Pensiri and S. Jitapunkul, "Speaker-Independent Thai Numerical Voice Recognition by using Dynamic Time Warping," Proceedings of the 18th Electrical Engineering Conference(EECON), pp.977-981, 1995.

[14] C. Wutiwiwatchai, S. Sae-Tung, V. Achariyakulporn, "Thai Language Speaker Identification System: Development Progress," NECTEC Technical Journal, Vol. II, No. 7, pp. 24-35, 2000.

[15] R. Boonsin, C. Jaruskulchai, "Thai Voice Command and Control for PocketPC,"[online] available: http://kucon.lib.ku.ac.th/Fulltext/KC4805003.pdf

[16] C.Wutiwiwatchai, V.Achariyakulporn, C. Tanprasert, "Text-dependent speaker identification using LPC and DTW for Thai language," Proceedings of the IEEE Region 10 Conference (TENCON 99), vol.1, pp.674-677. 1999.

[17] P. Tantanakit, "Automatic Continuous Speech Recognition of Thai Language using Hidden Markov Model," Master thesis, Prince of Songkla University, 2004.

[18] S. Jitapunkul, S. Luksaneeyanawin, V. Ahkuputra, E. Maneenoi, S. Kasuriya, P. Amornkul, "Recent Advances of Thai Speech Recognition in Thailand," IEEE conference APCCAS, pp.173-176, 1998.

[19] V. Ahkuputra, S. Jitapunkul, E. Maneenoi, S. Kasuriya, P. Amornkul, "Comparison of different techniques on Thai speech recognition," IEEE conference APCCAS, pp.177-180, 1998.

[20] J.Kenworthy, "Teaching English pronunciation," Longman House, London, 1987.

[21] NECTEC, "โครงการสื่อการสอนภาษามาลายูถิ่น [E-learning for local Malayu language,"[online]available: http://malayu.nectec.or.th/malayu_index.php

[22] S. Kasuriya, S. Kanokphara, N. Thatphitakkul, P. Cotsomrong, and T. Sunpethniyom, "Context-independent Acoustic Models for Thai Speech Recognition," ISCIT2004, pp. 991-994, Japan, October, 2004.

[23] M. Karnjanadecha, P. Kimsawad, and P. Tanthanakit, "HMM Based Speech Recognition of Continuous Thai Digits," Int. Symposium on Communications and Information Technology, pp. 271-274, Chiang Mai, Thailand, Nov. 14-16, 2001.

[24] M. Karnjanadecha, P. Kimsawad, W. Chukumnird, and K. Vaithayavanich, "An Automatic Speech Transcriber for the Thai Speech Corpus Project," Proceedings of the 3rd International Symposium on Communications and Information Technology, Vol. II, pp. 551-556, Songkhla, Thailand, Sept. 3-5, 2003.

[25] R. Thongprasirt, V. Sornlertlamvanich, P. Cotsomrong, S. Subevisai, and S. Kanokphara, "Progress Report on Corpus Development and Speech Technology in Thailand," Proc. of the Joint International Conference of SNLP Oriental COCOSDA 2002, pp. 300-306, Prachuapkirikhan, Thailand, 9-11 May, 2002.

[26] L. Tan and M. Karnjanadecha, "Modified Mel-Frequency Cepstrum Coefficient," Proceedings of the Information Engineering Postgraduate Workshop 2003, pp. 127-130, Songkhla, Thailand, Jan. 30-31, 2003.

[27] S. Suebvisai1, P. Charoenpornsawat, A. Black, M. Woszczyna, T. Schult, "Thai Automatic Speech Recognition," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2005, Vol.1, pp. 857-860. 2005.

[28] C. Wutiwiwatchai, S. Furui, "Thai speech processing technology: A review," Speech Communication Journal, Vol. 49, pp. 8-27, 2007.

[29] A. Chotimongkol, K. Saykhum, P. Chootrakool, N. Thatphithakkul, C. Wutiwiwatchai, "LOTUS-BN: A Thai Broadcast News Corpus and Its Research Applications," International Conference on Speech Database and Assessments 2009, pp. 44-50, 2009.

[30] F. Chelali, A. Djeradi, R. Djeradi, "Speaker Identification System based on PLP Coefficients and Artificial Neural Network," Proc. of the World Congress on Engineering 2011, July 6-8, London, Vol. II, 2011.

[31] P. Cotsomrong, T. Sunpetchniyom, S. Kasuriya, N. Thatphithakkul, C. Wutiwiwatchai, "LOTUS: Large vOcabulary Thai contInUous Speech Recognition Corpus," NSTDA Annual Conference S&T in Thailand: Towards the Molecular Economy (NAC2005), March 2005.

[32] A.Chotimongkol, N. Thatphithakkul, S. Purodakananda1, C.Wutiwiwatchai, P.Chootrakool, C. Hansakunbuntheung, A.Suchato, P. Boonpramuk, "The Development of a Large Thai Telephone Speech Corpus: LOTUS-Cell 2.0," Proc. of Oriental-COCOSDA '10, Kathmandu, Nepal, 2010.

[33] P. Tantanakit, and M. Karnjanadecha, "Phonetic Classification for Thai Speech Recognition," NCSEC2004 conference, Hadyai, Thailand, 20-22 October, 2004.

[34] H. Franco, L. Neumeyer, V. Digalakis, O. Ronen "Combination of machine scores for automatic grading of pronunciation quality," Speech Communication Journal, vol. 30 pp.121-130, 2000.

[35] V.Digalakis, H.Murveit, "GENONES: Optimizing degree of mixture tying in a large vocabulary hidden Markov model based speech recognizer," Proc. Internat. Acoust. Speech Signal Process, pp. 1537-1540, 1994.

[36] W.C. Huang, T.L. Chang-Chien, and H.P. Lin, "An Intelligent Multimedia E-Learning System for Pronunciations," Lecture Notes in Computer Science, vol. 4570/2007, pp. 84-93, 2007.

[37] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic Pronunciation Scoring for Language Instruction," Int. Conf. on ICASSP-97, vol. 2, pp. 1471-1474, 1997.

[38] S. Pakhomov, J. Richardson, M. Finholt-Daniel, and G. Sales, "Forced-Alignment and Edit-Distance Scoring for Vocabulary Tutoring Applications," Lecture Notes in Computer Science, Volume 5246/2008, pp. 443-450, 2008.

[39] N. Thatphithakkul, and B. Kruatrachue, "Denoise Speech Recognition Based on Wavelet Transform Using Threshold Estimation", EECON27 conference, November 2004.

[40] S. Suebvisai, P. Charoenpornsawat, A.W. Black, M. Woszczyna, T. Schultz, "Thai Automatic Speech Recognition," IEEE. Int. Conf. on ICASSP'05, pp. 857 – 860, 2005.

[41] N. Thatphithakkul, S. Kanokphara, "HMM Parameter Optimization Using Tabu Search," Proc. ISCIT2004, October 26-29, 2004.

[42] M. Karnjanadecha, and S. A. Zahorian, "Robust Feature Extraction for Alphabet Recognition," Proc. ICSLP 98, Sydney, Australia , vol. 2, pp. 337-340, 1998.

[43] M. Karnjanadecha, and S. A. Zahorian, "Signal Modeling for Isolated Word Recognition," Proc. ICASSP 99, vol. 1, pp.293-296, Phoenix , AZ., March 1999.

[44] M. Karnjanadecha, and P. Kimsawad, "A Comparison of Front-End Analyses for Thai Speech Recognition," Proc. ICSLP 2002, Denver , Colorado , USA., Sept.16-20, 2002.

[45] A. Deemagarn, A. Kawtrakul, "Thai connected digit speech recognition using Hidden Markov models," Int. Conf. on Speech and Computer SPECOM, pp. 731–735, 2004.

[46] A. Tungthangthum, "Tone recognition for Thai," Proc. APCCAS 1998, pp. 157-160, 1998.

[47] J.Chaiwongsai, W.Chiracharit, K.Chamnongthai, Y.Miyanaga, "An architecture of HMM-based isolated-word speech recognition with tone detection function," Int. Symposium on ISPACS 2008, pp.1-4, 2009.

[48] V. Ahkuputra, S. Jitapunkul, W. Ponrsukchandra, S. Luksaneeyanawin, "A speaker-independent Thai polysyllabic word recognition using Hidden Markov model," PACRIM conference, pp. 593–599.1997.

[49] S. Kanokphara, "Syllable Structure Based Phonetic Units for Context-Dependent Continuous Thai Speech Recognition," Proc. Eurospeech, pp. 797-800, 2003.

[50] A. Arya, "Interactive Face Animation - Comprehensive Environment (iFACE)," [online] Available: http://img.csit.carleton.ca/iface/

[51] N. Thubthong, B. Kijsirikul, "A syllable-based connected Thai digit speech recognition using neural network and duration modeling," Proc. ISPACS, Phuket. pp.785-788, 1999.

[52] N. Thubthong, B.Kijsirikul, A.Pusittrakul, "A method for isolated Thai tone recognition using combination of neural networks," Computational Intelligence,vol. 18 (3), pp. 313–335, 2001.

[53] C. Tanprasert, C. Wutiwiwatchai, S. Sae-Tnag, "Text-dependentspeaker identification using neural network on distinctive tone marks," Int. joint conf. IJCNN'99, vol. 5, pp. 2950-2953, 1999.

[54] E. Maneenoi, "Thai Vowel Phoneme Recognition Using Artificial Neural Networks," Chulalongkorn University, Bangkok, 1998.

[55] W. Pornsukchandra, S.Jitapunkul, V.Ahkuputra, "Speaker independent Thai numeral speech recognition using LPC and the back propagation neural network," Natural Language Processing Pacific Rim Symposium (NLPRS), pp. 585–588, 1997.

[56] V. Ahkuputra, S. Jitapunkul, and N. Jittiwarangkul, "A Comparison of Thai Speech Recognition Systems Using Hidden Markov Model, Neural Network and Fuzzy-Neural Network," Proc. ICSLP 98, Sidney, Australia, 1998.

[57] W.K. Lo, A.M. Harrison, H. Meng, L. Wang, "Decision Fusion for Improving Mispronunciation Detection using Language Transfer Knowledge and Phoneme-Dependent Pronunciation Scoring," Int. Symposium on Chinese Spoken Language Processing, pp. 1-4, 2008.

[58] F. Pan, Q. Zhao, Y. Yan, "Automatic Tone Assessment for Strongly Accented Mandarin," Proc. ICSP2006, Guilin, China, Vol. 1, pp. 758-761, 2006.

[59] Y. Liu, C. Yang, W. Ma, "Automatic Pronunciation Scoring for Mandarin Proficiency test based on Speech Recognition," Int. Symposium on Intelligent Ubiquitous Computing and Education, pp. 168 – 171, 2009.

[60] T. Masuko, T. Kobayashi, M. Tamura, J. Masubuchi, K. Tokuda, "Text-to-visual speech synthesis based on parameter generation from HMM," Proc. ICASSP1998, pp. 3745 – 3748, vol.6, 1998.

[61] C. Wutiwiwatchai, V. Achariyakulporn, S. Kasuriya, "Improvement of speaker verification for Thai language," Proc. EUROSPEECH, pp. 775–778, 2001.

[62] K.R. Aida–Zade, C. Ardil and S.S. Rustamov, "Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems," World Academy of Science, Engineering and Technology 19, pp. 74-80, 2006.

[63] N. Thatphithakkul, B. Kruatrachue, and C. Wutiwiwatchai, S. Marukatat, and V. Boonpiam, "Robust Speech Recognition Using KPCA-Based Noise Classification," ECTI Transactions on Computer and Information Technology, Vol.2(1), pp.45-53, May, 2006.

[64] PPA Innovation Co., Ltd., "Salika 2011," [online] Available: http://www.puttipan.com/salika

[65] NECTEC, "Welcome to Vaja," [online] Available: http://vaja.nectec.or.th/

[66] T. Chuensaichol, P. Kanongchaiyos, and C. Wutiwiwatchai, "Thai Lip-sync : Mapping Lip Movement to Thai Speech," Wisawakammasat Journal, Vol. 3, No. 2, pp.33-42, 2011.

[67] R. Parent, "Computer Animation, Second Edition: Algorithms and Techniques," 2nd edition, Morgan Kaufmann, 2007.

[68] M. M. Cohen and D. W. Massaro, "Modelmg Coarticulation in Synthetic Visual Speech," M. Thalmann & D. Thalmann (Eds.) Computer Animation '93. Springer Verlag, Tokyo.

[69] J. Schroeter, J. Ostermann, H. P. Graf, M. Beutnagel, E. Cosatto, A. Syrdal, A. Conkie, and Y. Stylianon, "Multimodal speech synthesis," Proc. ICME 2000, Vol. 1, pp. 571-574, 2000.

[70] T. Sy-Sen, A.W.C. Liew, Y. Hong, "Lip-sync in human face animation based on video analysis and spline models," Proceedings. 10th International on Multimedia Modelling Conference, pp. 102-108, 2004.

[71] K. Balci, E. Not, M. Zancanaro, and F. Pianesi, "Xface open source project and smil-agent scripting language for creating and animating embodied conversational agents," Proc. ACM Multimedia 2007, Germany, pp. 1013-1016, 2007.