An Investigation on Speech Perception Under Effects of Coarticulation

Trung-Nghia Phung, Mai Chi Luong, and Masato Akagi

***Abstract*—In this paper, we investigated the speech perception under effects of coarticulation within syllables, including estimating the nuclei and coarticulated transition intervals of phonemes, and investigating the perception of nuclei and coarticulated transition intervals of phonemes within syllables. To complete these objectives, we firstly proposed the folded spectral transition measure to estimate the boundary points between the nuclei interval and coarticulated transition intervals. Then, we conducted experiments to investigate the syllable identification and syllable quality evaluation under effects of coarticulation. The experimental results show that human can identify the syllables with the retaining intervals equal or wider the coarticulated transition intervals. Besides, the experimental results show that the nuclei intervals are still important for speech quality perception; the quality of truncated syllables retaining the two phoneme nucleus and coarticulated transition interval between the two phonemes within syllable is equivalent with the quality of original two-phoneme syllables.**

***Index Terms*—Coarticulation, speech perception, spectral transition measure.**

I. Introduction

Coarticulation is a phonological phenomenon, always occurring in all languages for all sequences of sounds not separated by pauses. Without appropriate coarticulation the resulting speech sounds unnatural and is hard to understand.

In the literature, many coarticulation models have been proposed [1]-[3]. Applying these models, speech perception under effects of coarticulation has been investigated. Furui [4] showed that human even can identify the vowel, consonant and syllable only with the transitions when vowel nucleus are truncated. However, state of the art speech recognition and concatenative speech synthesis show that the vowel nucleus, related to static information of syllables, play important role. This contrary needs to be clarified by further experiments.

Spectral transition is closely related to speech perception in coarticulation [4]. Based on spectral transition measure (STM), there are some methods to estimate the boundary points between the phonemes [4], [5] and to estimate the nuclei points, related to the locations of the idealized articulatory targets of phonemes [6], [7].

However, it is still a lack of efficient methods for automatically estimating the nuclei intervals and the coarticulated transition intervals of the phonemes in each syllable.

This paper therefore is to investigate the speech perception under effects of coarticulation within syllables, including estimating the nuclei and coarticulated transition intervals of phonemes, investigating the perception of nuclei and coarticulated transition intervals of phonemes within syllables. To complete these objectives, we conducted experiments in a Vietnamese speech dataset, presented in the section V.

Manuscript received May10, 2012; revised June 12, 2012.

Trung-Nghia Phung and Masato Akagi are with the Japan Advanced Institute of Science and Technology, Ishikawa, Japan (e-mail: ptnghia@jaist.ac.jp, akagi@jaist.ac.jp).

Mai Chi Luong is with the Vietnam Institute of Information Technology, Hanoi, Vietnam (e-mail: lcmai@ioit. ac.vn).

II. Coarticulation Theory and Models

The articulatory phonological theory [8] regards gestures as the basic units of phonological contrast. A gesture is the movement of one articulator from an articulatory position characteristic of one speech sound to an articulatory position characteristic of the next speech sound. Articulatory gestures overlap. Articulatory overlap is the basis of coarticulation.

In the most basic model of articulatory planning, Locus [1], each phoneme has a single ideal articulatory target for each contrastive articulator independent of the neighboring phonemes. These ideal targets locate around the centers of the phoneme nucleus [6], [7]. Under effects of coarticulation, the transition between two phonemes is described as the movement between the two ideal targets of the phonemes. This transition shares the articulatory and acoustic characteristics of the two targets of both phonemes and gradually changes from being predominantly like the first phoneme target to predominantly like the second phoneme target.

The Kozhevnikov-Chistovich model [2] founds coarticulation within syllable but not across syllables. Thus, they have been used to model coarticulation between phonemes within syllables. These models are considered suitable for modeling speech in monosyllable languages, in which coarticulation is supposed to occur between phonemes within syllable rather than across the syllables.

The Wickelgren [3] is the model that mentally codes speech units as context-sensitive units, thus each phoneme is just affected by the two nearest neighboring phonemes. Wickelgren is the most popular model currently used in HMM-based speech recognition and synthesis.

Target undershoot is a phonological phenomenon, occurring when there is insufficient time for an articulator to reach its target position. Thus, target undershoot is not clear described by simple models of articulatory planning. However we can avoid the target undershoot by using the speech that is pronounced clearly and slowly.

In articulatory phonology theory of Browman [8], there is

more than one single target in each phoneme. Targets of one phoneme might be located at different locations in time. The articulatory phonology theory suggests us a supposition that it exists a nuclei interval of each phoneme. In this nuclei interval, there are some intra-targets of the phoneme, coarticulation occurs between these targets within the nuclei interval. The phoneme to phoneme coarticulation only occurs from the right outermost target of the prior phoneme to the left outermost target of the next phoneme. The coarticulated transition interval between phonemes therefore is approximately estimated by the transition interval between the two outermost targets. It is known that there is no articulatory target at the transition interval; thus, we attempted to estimate the phoneme nuclei interval separately with the coarticulated transition interval of each phoneme in this work.

III. Speech Perception Under Effects of Coarticulation

In the literature, many researchers have taken into account the effects of coarticulation on speech perception. Furui [4] found that the dynamic spectral information caused by coarticulation is critical for vowel, consonant and syllable perception. Furui also showed that human even can identify the vowel, consonant and syllable only with the transition when vowel nuclei is attenuated to silence or truncated.

State of the art computer-based speech recognition show that both static (eg. MFCC) and dynamic (eg. Delta MFCC) information are important [9]. Speech could not be perfect recognized only with dynamic features. Moreover, in state of the art concatenative speech synthesis, the vowel nucleus play critical role for high quality speech synthesis [10], [11].

Therefore, although human can recognize speech with only dynamic features, both of static (eg. phoneme nucleus) and dynamic features (eg. transitions) are important for perfectly recognizing and naturally perceiving continuous speech. This supposition has not been clarified in the previous researches and need to be confirmed by further experiments.

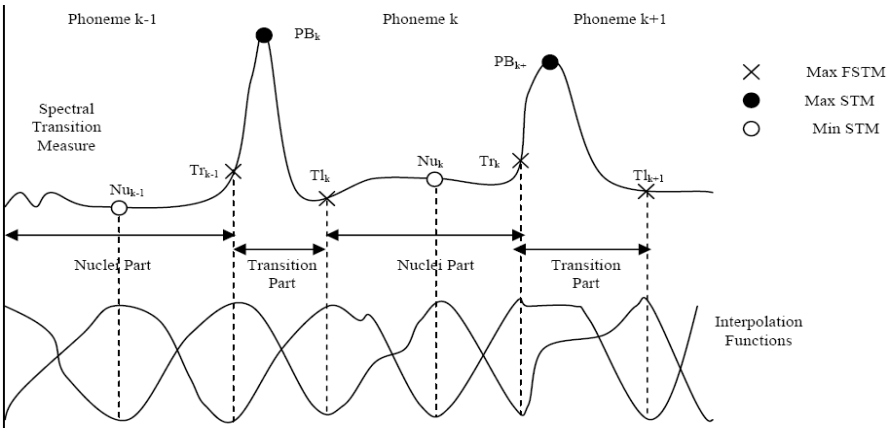

Fig. 1. Transition and nuclei intervals

In experiments of Furui [4], the coarticulated transition interval, bearing the most important information of syllables, was estimated manually with an experimental duration length. The experimental duration length much depends on the specific conditions and could not be widely applied for speech applications.

Current concatenative speech syntheses are only high quality if we have enough data of all phonetic contexts, such as in unit selection synthesis. If we have limited data and speech might need to be synthesized by replacing phoneme from different contexts, we need to consider coarticulation effects and modify the coarticulated transition interval to fit with the required context. However, current concatenative speech syntheses only averagely smooth the boundary points instead of the whole coarticulated transition interval between the neighboring phonemes [10], [11]. It causes the quality of concatenative speech syntheses with limited data is not high.

The above analyses lead us to a problem that how to automatically estimate the coarticulated transition interval between neighboring phonemes within syllables. The answer for this problem can be solved by using the proposed folded transition measure in this paper.

IV. Spectral Transition Measure and Folded Spectral Transition Measure

A. Spectral Transition Measure

Transient portions of the spectrum of speech signal, related to dynamic spectral information, have been considered important for speech perception [4]. Furui [4] defined the spectral transition measure (STM) for representing the magnitude of spectral dynamic.

The STM at the time *t*, *D(t)*, is computed as

$$D(t) = (\sum_{i=1}^{p} a_i^2) / p \quad (1)$$

$$\text{where } a_i = (\sum_{n=-n0}^{n0} C_i(n).n) / (\sum_{n=-n0}^{n0} n^2) \quad (2)$$

Here $C_i(n)$ is the i^{th} spectral coefficient ($1 \le i \le p$) at the n^{th} frame within an interval whose center is the time t, and $-n0 \le n \le n0$.

The regression coefficient a_i, corresponds to the linear variation of the spectral envelope pattern in a unit time. Consequently, $D(t)$, which is the mean-square value of a_i, $i = 1..p$, corresponds to the variation of the smoothed spectral envelope.

Many researches [4, 5] show that the maximum of STM can be approximated as the boundary of the phonemes. Besides, the minimum of STM can be considered the center of phoneme nuclei, and approximated as location of idealized articulatory target [6, 7].

B. Folded Spectral Transition Measure

Denote the center point of phoneme k is Nu_k, the boundary point of phoneme $k\text{-}1^{th}$ and k^{th} is PB_k. The phoneme k^{th} is estimated in the interval from PB_k to PB_{k+1}.

The FSTM is geometrically defined as a relatively changing rates of STM.

$$FSTM = \begin{cases} \log(\Delta_{t+1}/\Delta_t), if\ Nu_{k-1} < t < PB_k \\ \log(\Delta_t/\Delta_{t+1}), if\ \ PB_k < t < Nu_k \end{cases} \quad (3)$$

$$\text{where } \Delta_t = D(t) - D(t-1) \quad (4)$$

and $D(t)$ is the STM at the time t.

For each phoneme k^{th}, there are two folded transition points at the two sides of the center point Nu_k, Tr_k at the right side and Tl_k at the left side. Tr_k and Tl_k is defined as the maximum of FSTM as shown in Fig.1.

We can observe that the maximum of the FSTM estimates the points where the transition measure relatively changing rates are suddenly changed; those are the *folded transition points*. We proposed to estimate the coarticulated transition interval between the phoneme $k\text{-}1^{th}$ and k^{th} as the interval between the Tr_{k-1} and Tl_k, shown in Fig.1. The proposed estimation is based on the supposition that when changing from stable to dynamic region (and in inverse case), the relatively changing rate is suddenly increased (decreased) at the onset of dynamic (stable) region.

In the experiment in section V.C, we will investigate whether the folded transition points approximate the perceptual critical points of Furui's experiments [4], and confirm that it can be used to estimate the coarticulated transition interval between phonemes.

V. Experiments

A. Data Preparation

In this work, we conducted some experiments to investigate the coarticulation in Vietnamese. We based on the Kozhevnikov-Chistovich and Wickelgren models to design the stimuli for experiments.

Vietnamese is a mono-syllable language and coarticulation occurs between the neighboring phonemes within syllable rather than across syllables [12]. Based on the Kozhevnikov-Chistovich model, we designed the stimuli to investigate the coarticulation between neighboring phonemes within each syllable and ignored the coarticulation across the syllables. To achieve this goal, we chose a dataset of discrete Vietnamese syllables, in which each syllable was pronounced alone to remove the effects of coarticulation between syllables. We also chose the syllables which were pronounced clearly and slowly to avoid target undershoot in coarticulated phonemes within syllables.

Based on Wickelgren model, we designed the stimuli to investigate the coarticulation between only the two nearest neighboring phonemes. Thus, we chose the two-phoneme syllable set, which are three groups CV, VC and VV. Each group had 10 syllables. These syllables were sampled at 11025 Hz, quantized to 16 bits in single channel mode. The chosen dataset does not cover all Vietnamese syllable structures but cover three popular kinds of Vietnamese syllable structures, CV, VC and VV.

B. Stimuli and Method for Experiments

In order to confirm the duration of the dynamic, refered to as the coarticulated transition interval, Furui [4] conducted an experiment to investigate the identification scores of truncated syllables as a function of truncation positions. The beginnings and endings of the syllable were jointly truncated. The perceptual critical point was estimated as the truncation position when identification scores rapidly changed from a low score to a very high score exceeded about 80%. The interval between the two neigboring perceptual critical points, located at two sides of a phonemes boundary point, was considered the coarticulated transition interval between the two neigboring phonemes within a syllable.

The folded transition points in this paper, used to estimate the coarticulated transition intervals within syllables, therefore are refered to as the perceptual critical points that Furui mentioned. We will confirm whether the identification scores rapidly changed from a low score to a very high score when the truncation passing the folded transition points.

The stimuli of the experiments in this paper were developed from those of Furui, where the beginnings and endings of the syllable were jointly truncated by fixed size truncation steps. The spectral feature for STM and FSTM used in this paper is the line spectral frequency (LSF). Because the formant transition are considered corresponding with transition of articulators, the reason for selecting LSFs in this paper is that LSFs parameters are closely related to formant frequencies, and they can be estimated reliably.

Perceptual linear prediction (PLP) was confirmed to be more robust with environments than original linear prediction (LP) [13]. We expected that it is also robust with vowel neutralization phenomenon, caused by coarticulation. Thus, we used a perceptual LSF computed by PLP, called PLP-LSF, instead of using LSF computed by original linear prediction parameters.

The number of LSF coefficients p was *15* for each overlapped frame, where the frame size and interval were *10 ms* and *3 ms* in turn. The frame size and interval were set small to detect the target points more locally. The time shift *n0* in STM was set to *2* frames. We truncated the syllable around the folded transition point, where truncation step index is set to be zero.

The truncation step was *5 ms*, smaller than the step size used in Furui's experiments [4] to cope with the short phonemes.

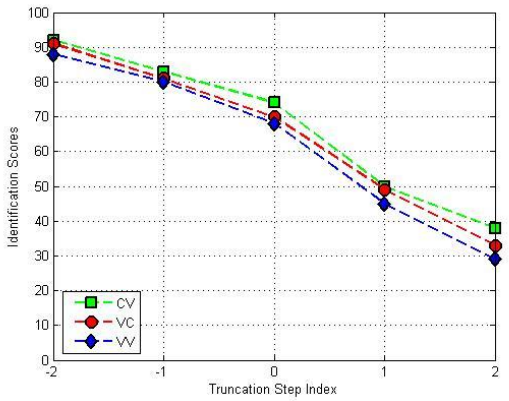

Fig. 2. Identification Scores for Truncated Syllables

Syllables were concurrently truncated in both sides left and right of the coarticulated transition interval. After being truncated, the speech waveform amplitude was linear attenuated to reach zero at the truncation point.

We chose a group of five Vietnamese native listeners for these experiments. The listeners were required to identify the syllables and evaluate the quality of the syllables.

C. Experiment on Syllable Identification under effects of coarticulation

In this experiment, we will confirm that the perceptual critical points in the experiments of Furui [4] can be approximated as the folded transition points *Tr* and *Tl*, which can be automatically computed by the algorithm presented in the sub section IV.B. The 30 original syllables and their truncated versions were played in random order to the listeners. The listeners had to identify the syllables by writing the name of the syllables. The number of syllables correctly recognized by listeners was taken as the identification score.

The result is shown in Fig.2, where syllables were truncated for both sides around the folded transition points, the step index of truncation equals *0* means the truncated syllable is the same region between two folded transition points, equal -1 means the truncated syllable is the region between two folded transition points added by two adjacent frames in both sides, equal 1 means the truncated syllable is the region between two folded transition points subtracted by two outermost frames, and so on.

The results show that the identification scores for the retaining intervals that equal or wider the coarticulated transition interval, corresponding with truncation step indexes smaller than or equal as zero, were very high. It confirms that human can identify the syllables only with the coarticulated transition interval. The identification scores for truncated intervals, corresponding with truncation step indexes larger than zero, are very low. It confirms that the two points *Tl* and *Tr* were approximate the perceptual critical points in the Furui's experiments [4]. The greater coarticulation effects, the more important information remain in the coarticulated transition intervals, then the higher scores human can identify the syllables.

In our experiments, we also found that coarticulation in Vietnamese VV structure is weaker than in CV and VC structures, thus the identification scores for truncated VV syllables under effects of coarticulation was worst.

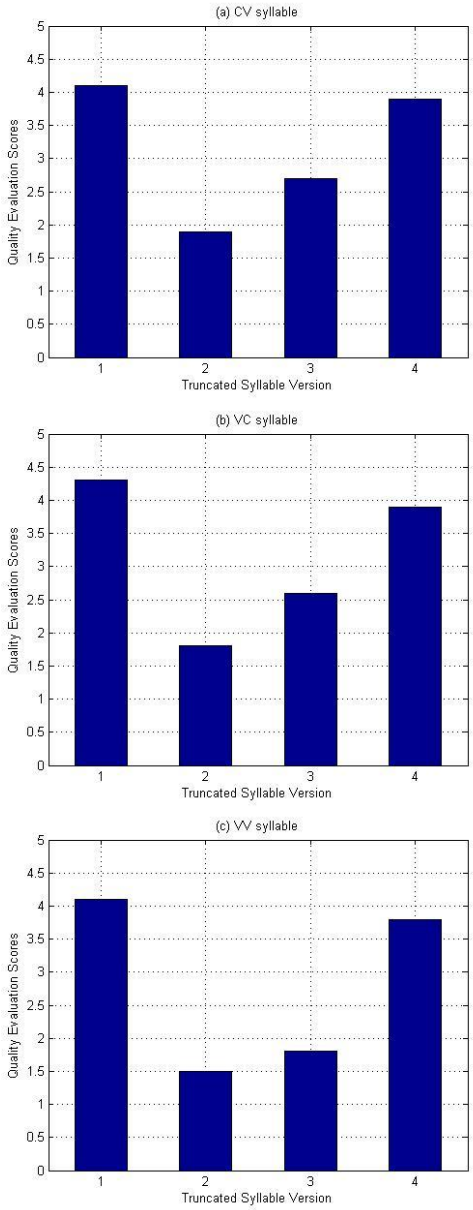

Fig. 3. Quality evaluation scores for truncated syllables

D. Experiment on Evaluating Quality of Syllable under effects of coarticulation

In this experiment, we will confirm that although human can identify the syllable only with coarticulated transition interval, the nuclei interval is still very important for speech quality perception. We provided the four versions of the syllables for listeners, including the original syllables (version 1); the truncated syllables retaining only coarticulated transition interval , as the interval between Tr_{k-1} and Tl_k in Fig. 1 (version 2); the truncated syllables retaining the interval between nuclei points of the two phonemes in syllables, as the interval between Nu_{k-1} and Nu_k in Fig.1 (version 3); and the truncated syllables retaining both of the two nuclei intervals and the coarticulated transition interval between the two phonemes, as the interval between Tl_{k-1} to Tr_k in Fig.1 (version 4).

The listeners had known the writing characters of syllables before, and were required to evaluate the quality of syllables on a five-point scale (1: bad, 2: poor, 3: fair, 4: good, 5: excellent). The result is shown in Fig.3, in which we see that the quality of truncated syllables retaining only coarticulated

transition interval is worst, the quality of truncated syllables retaining the two nuclei intervals and the coarticulated transition interval between the two phonemes is equivalent with that of original syllables. We also can observe that human perceived the quality of original syllables CV, VV and VC almost the same. But human perceived the quality of truncated syllables retaining only coarticulated transition interval of CV, VC better than that of VV syllables.

VI. Conclusion

In this paper, we investigated the speech perception under effects of coarticulation within syllables, including estimating the nuclei and coarticulated transition interval of phonemes, investigating the perception of nuclei and coarticulated transition interval of phoneme within syllable.

The folded transition points estimated by maxima of FSTM were confirmed to be approximate with the perceptual critical points in experiments of Furui [4]. Thus, they can be used to estimate the boundary points of nuclei and transition intervals within syllables.

Two experiments to investigate the syllable identification and syllable quality evaluation under effects of coarticulation were also conducted. The experimental results show that human are able to identify the syllables only with coarticulated transition interval but the nuclei interval is still important for speech quality perception.

References

[1] P. Delattre, "Coarticulation and the Locus Theory," *Studia Linguistica*, vol. 23, no. 1, pp. 1–26, June 1969.

[2] V. A. Kozhevnikov and L. A. Chistovich, "Rech: Artikulatsiya i Vospriatatie (Moscow-Lenningrad)," *Trans. Speech: Articulation and Perception. Washington, DC: Joint Publication Research Service*, no. 30, pp. 543, 1965.

[3] W. A. Wickelgren, "Context-sensitive coding, associative memory, and serial order in speech behavior," *Psychological Review*, vol. 76, pp. 1-15, 1969.

[4] S. Furui, "On the role of spectral transition for speech perception," *J Acoust Soc Am*, 80(4), pp. 1016-25, 1986.

[5] S. Dusan and L. Rabiner, "On the relation between maximum spectral transition positions and phone boundaries," *Proc. 9th Int. Conf. Spoken Lang. Process*, pp. 17-21, 2000.

[6] P. C. Nguyen, T. Ochi, and M. Akagi, "Modified Restricted Temporal Decomposition and Its Application to Low Rate Speech Coding," *IEICE Trans. Inf. and Syst.*, vol. E86–D, no. 3, 2003.

[7] A. C. R. Nandasena, P. C. Nguyen, and M. Akagi, "Spectral stability based event localizing temporal decomposition," *Computer Speech and Language 15*, pp. 381–401, 2001.

[8] C. P. Browman and L. Goldstein, "Articulatory phonology: an overview," *Phonetica*, vol. 49, pp. 3-4, pp. 155-180, 1992.

[9] L. Rabiner, "Fundamentals of Speech Recognition," Prentice Hall *Signal Processing Series*, 1992.

[10] A. J. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *ICASSP*, vol. 1, pp. 373 – 376, 1996.

[11] J. Wouters, "Control of Spectral Dynamics in Concatenative Speech Synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 30-38, January 2001.

[12] Đ. T. Thuật, N. Â. T. Việt, and N. X. B. Đ. H. Q. Gia, 2003.

[13] H. Hemamsky, "Perceptual Linear Predictive (PLP) analysis of speech," *J* Acoust *Soc Am*, vol. 87, no. 4, pp. 1738 – 1752, 1990.

[14] J. K. Lee, "The asymmetry of C/V coarticulation in CV and VC structures and its implications in phonology," *Studies in the Linguistic* Sciences, vol. 27, no. 1, pp. 139-152, 1997.

Trung-Nghia Phung received his B.E. in Electronic & Telecommunication engineering from the Hanoi University of Technology in 2002 and his M.E. in Electronic & Telecommunication engineering from the Vietnam National University, Hanoi, in 2007.

He has been with the Thai Nguyen University of Information and Communication Technology from 2003. He has also been a Ph.D. candidate at the School of Information Science of the Japan Advanced Institute of Science and Technology (JAIST) since 2009.

His research interests are speech and biomedical signal processing.

Mai Chi Luong received her B.E. from Faculty of Applied Mathematics, Kishnov Unversity (Soviet Union former) in 1981, and her Ph.D. from the Institute of Information Technology (IOIT) in 1991.

She joined the Laboratory of Pattern Recognition, IOIT in 1982. From 1982 to present, she is working as a senior researcher and got her Associate Professsor in 2005. During 1987-1990, she was associated with the International Basic Laboratory on Intelligence Artificiel, Institute of Cybernetics, Bratislava under Slavak Academy of Science as a visiting fellow.

Her research interests include pattern recognition, machine learning, speech recognition and synthesis.

She is a member of the Institute of Electrical and Electronic Engineering (IEEE). Dr. Luong Chi Mai received the Kovalevskaia Award for her outstanding contribution on the R&D in Vietnam for 2010.

Masato Akagi received his B.E. from Nagoya Institute of Technology in 1979, and his M.E. and Ph.D. Eng. from the Tokyo Institute of Technology in 1981 and 1984.

He joined the Electrical Communication Laboratories of Nippon Telegraph and Telephone Corporation (NTT) in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992 he has been on the faculty of the School of Information Science of the Japan Advanced Institute of Science and Technology (JAIST) and is now a full professor.

His research interests include speech perception, the modeling of speech perception mechanisms in human beings, and the signal processing of speech. During 1998, he was associated with the Research Laboratories of Electronics at MIT as a visiting researcher, and in 1993 he studied at the Institute of Phonetics Science at the University of Amsterdam.

He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Acoustical Society of Japan (ASJ), the Institute of Electrical and Electronic Engineering (IEEE), the Acoustical Society of America (ASA), and the International Speech Communication Association (ISCA).

Dr. Akagi received the IEICE Excellent Paper Award from the IEICE in 1987, the Best Paper Award from the Research Institute of Signal Processing in 2009, and the Sato Prize for Outstanding Papers from the ASJ in 1998, 2005, 2010 and 2011. He is the current president of the ASJ.