An Investigation on Perceptual Line Spectral Frequency (PLP-LSF) Target Stability against the Vowel Neutralization Phenomenon

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

***Abstract*—Coarticulation is a phonological phenomenon, always occurring in all sequences of sounds not separated by pauses. Analyses in coarticulation of speech reveal that articulation targets are incomplete in neutralized sound, and it is difficult to estimate incomplete articulatory targets of phonemes due to their sensitivity. In this paper, we firstly proposed a acoustical model of coarticulation of phonemes within syllables. After that, we investigated the stability of spectral targets affected by the vowel neutralization phenomenon. The experimental results show that the proposed coarticulation model decomposed speech into context-insensitive event targets, which are close with articulatory targets, and the context-sensitive event functions, which closely represent the movements between the adjacent targets. In addition, the PLP-LSF was shown as a stable spectral target against vowel neutralization phenomenon in our proposed coarticulation model.**

***Index Terms*—Coarticulation, Neutralization, Perceptual Linear Prediction, Line Spectral Frequency, Temporal Decomposition.**

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, 2, 3, 4].

In the most basic model of articulatory, Locus [1], each phoneme has a single ideal articulatory target for each contrastive articulator independent of the neighboring phonemes. 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. This model is considered suitable for modeling speech in monosyllable language, 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 with a supposition that each phoneme is just affected by the two nearest neighboring phonemes.

Extended from basic of Locus theory, the articulatory phonology theory of Browman [4] shows that there is more than one single target in each phoneme. Targets of one phoneme might be located at different locations in time.

Although there are many coarticulation models have been proposed. There is still a lack of simple models, which are easy to be implemented in speech applications, and directly performed with acoustic data. In this research, we used the spectral transition measure (STM) [5], the folded STM (FSTM), and the temporal decomposition (TD) [6, 7] to model the coarticulation between intra-targets within nuclei intervals of phonemes, as well coarticulation between inter-targets of neighboring phonemes. The details of the proposed model are presented in section IV, and the experimental results are presented in the section VI.

Using the STM, the boundary points between the phonemes and the nuclei points, related to the locations of the idealized articulatory targets of phonemes, could be estimated [5, 6, 7]. The nuclei intervals, containing static spectral targets, and the transition intervals, containing spectral dynamics, could be also manually estimated [5]. While speech dynamics, known to be context sensitive, bear a lot of phonetic information of speech, thus they are very important to speech intelligibility [5]. The static spectral targets, bear both linguistic/phonetic information and non/paralinguistic information of speech, are very important to both speech intelligibility and quality [8]. Therefore, static spectral targets are usually required as stable as possible for reliable recognition and synthesis tasks. However, analyses in coarticulation of speech reveal that articulation targets are incomplete in neutralized sound, and the spectral targets of neutralized phonemes are not stable and not easy to be estimated due to their sensitivity. In this paper, we investigated the stability of spectral targets of vowel nucleus under effects of the vowel neutralization phenomenon in many different phonetic contexts. The details are present in section V, VI.

II. Spectral Transition Measure And Folded Spectral Transition Measure

A. Spectral Transition Measure

The STM at the time *t*, *STM(t)*, wass defined [4] as

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

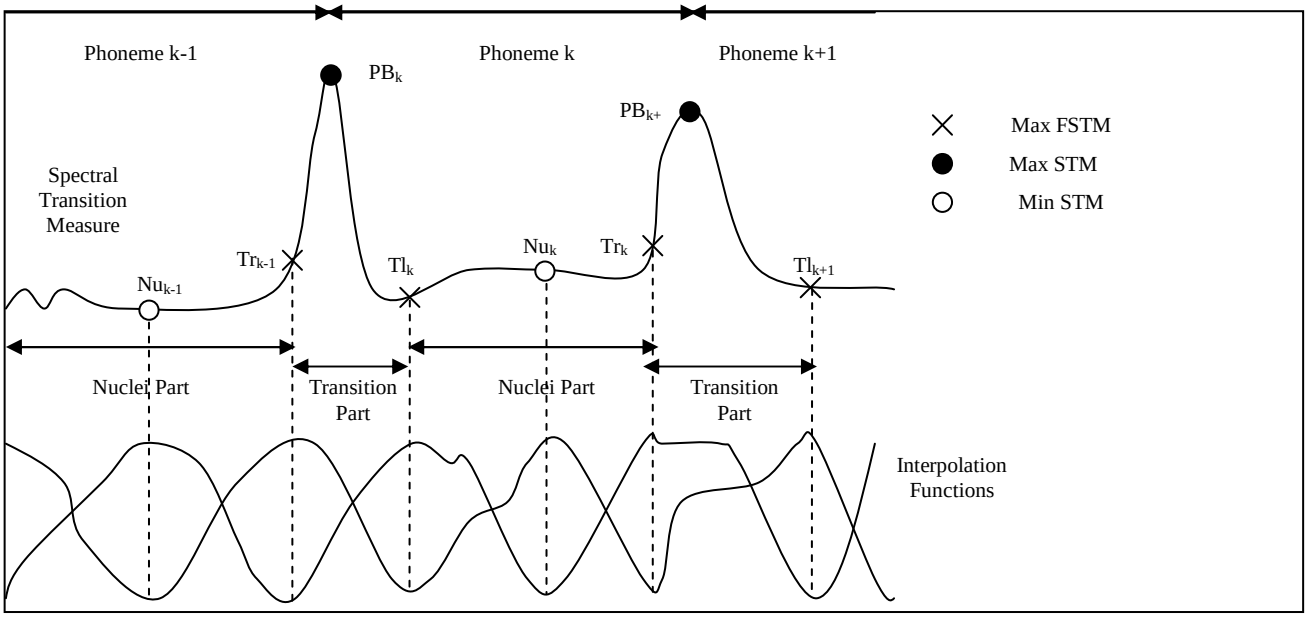

Fig. 1. Transition and Nuclei Intervals

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

Here $C_i(n)$ is the i^{th} order 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, STM(t), which is the mean-square value of a_i, $i = 1..p$, corresponds to the variation of the smoothed spectral envelope.

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

B. Folded Spectral Transition Measure

Denote the center point (min STM) of phoneme k is Nu_k, the boundary point of phoneme $(k-1)^{th}$ and k^{th} is PB_k. The phoneme k^{th} is determined in the interval from PB_k to PB_{k+1}.

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

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

where $\Delta_t = D(t) - D(t-1)$ *(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. In this paper, we proposed to estimate the coarticulated transition interval between the phoneme $(k-1)^{th}$ 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 region.

III. TEMPORAL DECOMPOSITION

TD [6, 7] yields a linear interpolation of a time sequence of spectral parameters in terms of a series of time-overlapping event functions and an associated series of event vectors as given in Eq. (5).

$$\hat{y}(n) = \sum_{k=1}^{K} a_k \phi_k(n), 1 \le n \le N \quad (5)$$

where a_k and $\phi_k(n)$ are the k^{th} event vector and k^{th} event function, respectively. $\hat{y}(n)$ is the approximation of $y(n)$, the n^{th} spectral parameter vector, produced by the TD model. The second order TD model used in [5], where only two adjacent event functions overlap, is given in Eq. (6).

$$\hat{y}(n) = a_k \phi_k(n) + a_{k+1}\phi_{k+1}(n), n_k \le n \le n_{k+1} \quad (6)$$

where n_k and n_{k+1} are the locations of event k^{th} and event $(k+1)^{th}$, respectively. The restricted second order TD (RTD) model was utilized in [6] with an additional restriction to the event functions in the second order TD model that all event functions at any time sum up to one. Eq. (6) can be rewritten as

$$\hat{y}(n) = a_k \phi_k(n) + a_{k+1}(1 - \phi_k(n)), n_k \le n \le n_{k+1} \quad (7)$$

A modification of RTD called modified RTD (MRTD), using line spectral frequency (LSF) parameter, was proposed [6]. In this work, we also used the MRTD because of its compactness and efficiency, and we proposed a method for modeling the coarticulation in syllable using MRTD.

IV. PROPOSED COARTICULATION MODEL

The Locus model shows that the coarticulation between two targets is described as the transition movement between the two neighboring targets. The Kozhevnikov-Chistovich model founds coarticulation within syllable rather than across syllables. The Wickelgren model shows that each target is just affected by the two nearest neighboring targets. Extended from Locus theory, the articulatory phonology theory of Browman shows that there is more than one single target,

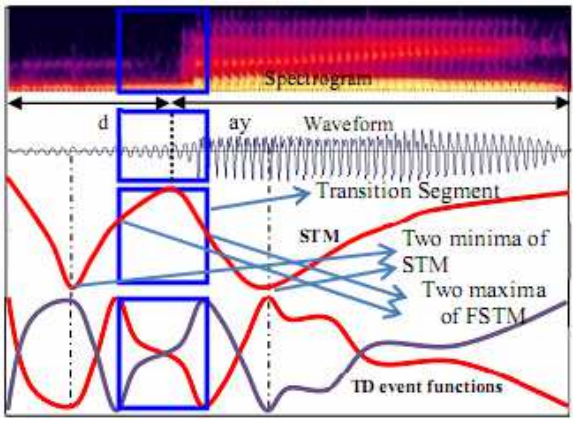

Fig. 2. Modeling the Coarticulation by TD Event Functions

located at different locations in time, in each phoneme. Theory of Browman 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 coarticulated transition, referred to as inter-targets transition, 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.

Based on basic models and theories of Locus, Kozhevnikov-Chistovich, Wickelgren, and Browman, as well our previous analysis, we proposed a coarticulation model for monosyllable languages, performed directly with acoustic data, presented coarticulation as the transition movements between adjacent static targets of phonemes within a syllable.

The proposed coarticulation model is described in Fig. 2.

In this proposed coarticulation model, STM and FSTM were used to estimate the MRTD event locations, referred to as context-insensitive target locations, in which two outermost targets of each phonetic unit are approximately located at the onset and offset of the coarticulated transition interval at the two sides. The coarticulated transition interval of two adjacent phonemes, describing the contextual effects of phonemes within a syllable, is then represented by the interpolation region, modeled by two overlapped TD event functions, between these two outermost targets of these two phonemes. The interpolation performance of the proposed model was evaluated and presented in the subsection VI.B.

Using the proposed model, we separated the context-insensitive static features, related to TD event targets, and context-sensitive dynamic features, related to TD event functions in continuous speech under effects of coarticulation. The context-sensitive transition movements between neighboring phonemes within syllable were described be two overlapped event functions, thus it could be modified to fit with a new context. The static context-insensitive event targets, representing the context-independent characteristics of phonemes, were expected to be stable and reliable. Therefore, we considered and investigated the stabilities of spectral targets under effects of coarticulation, as presented in section V and subsections VI.C, VI.D.

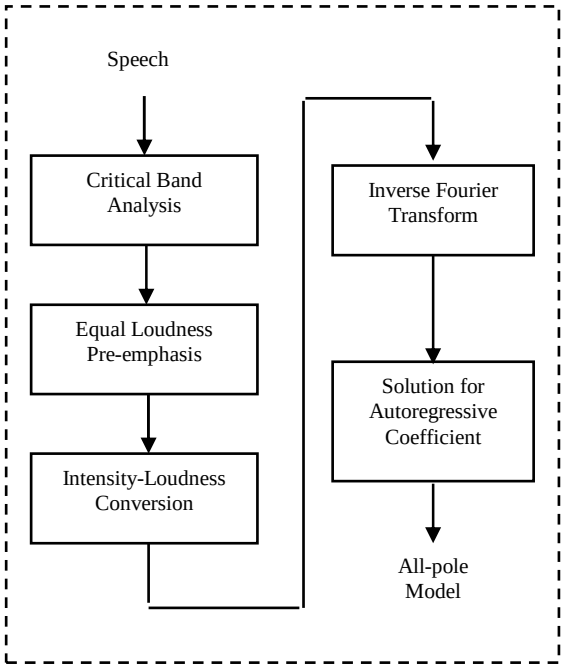

Fig. 3. Perceptual Linear Predictive (PLP)

V. STABILITY OF SPECTRAL TARGETS

Spectral targets estimated by the proposed coarticulation model were expected to be a stable context-insensitive phonetic parameter. However, in some cases, articulation targets are sensitive. Analyses in coarticulation of speech reveal that articulation targets are incomplete in neutralized sound. Human can perceive each phoneme neutralized by coarticulation nearly as if it were uttered clearly without neutralization. However, computer algorithms have not worked well to identify incomplete targets because of their sensitivities. To represent spectral target close with human hearing therefore is expected to improve the stability of the spectral target against the vowel neutralization phenomenon.

LSF is closely related to formant frequency, considered corresponding with transition of articulators; LSF can be also reliably estimated. Therefore, the spectral features using in our study are LSF and its variants. The original LSF is computed from the original linear prediction coefficient (LPC). However, it has been showed that LPC is not environmentally robust. Human can perceive speech even in highly noisy environment, and to represent the spectral parameters close with human hearing can improve the noise robustness. Perceptual linear predictive (PLP) [9] and RASTA [10], built closely with human hearing, were proposed. The PLP was built based on three techniques, the critical band spectral resolution (bark-scale), equal-loudness pre-emphasis and intensity-loudness power law [9]. The RASTA was an improvement of PLP which makes PLP more robust to linear spectral distortions [10]. Diagram of PLP is presented in Fig.3. PLPs, including PLP and RASTA, have been considered the robust spectral representation in noisy environments. We expected that combinations of LSF and PLP including Bark-Scale LSF, PLP-LSF, and RASTA-LSF, can improve the stability of spectral targets. Therefore, we investigated their stabilities under effects of coarticulation. The investigation results are shown in subsection VI.C, VI.D.

VI. EXPERIMENTS AND EVALUATIONS

A. Data Preparation

The speech corpus used in the experiments is DEMEN567, also called Vnspeech corpus [11], built in the Institute of Information Technology of Vietnam (IOIT). Speech was originally sampled at 11025 Hz, and re-sampled at 8 KHz, quantized to 16 bits, single channel.

In the first experiments to evaluate interpolation performance of the proposed coarticulation model, the dataset consisted of 50 utterances, extracted from DEMEN567.

In the second experiment to evaluate the effects of neutralization on spectral target stability, we used a syllable set with structure CVC, extracted from DEMEN567, and divided into 13 groups corresponding to 13 Vietnamese basic vowels. To clearly evaluate the effects of neutralization in spectral target stability, we chose the CVC syllables spoken with fast speaking rate.

B. Evaluating Interpolation Performance of the Proposed Coarticulation Model

In this section, we conducted the experiment to confirm that the TD event locations, chosen by STM and FSTM in our proposed model could be approximate with static spectral target locations and could improve the interpolation performance.

We interpolated syllables by two methods for comparison. In the baseline method, the events were located in positions with equally spaces. In proposed method, the events were located at the Tr, Tl, which are the maxima of FSTM, and Nu(s), which are local minima of STM, with the equivalent number of events. Log spectral distortion (LSD) was used to evaluate the interpolation performance of the MRTD with proposed event locations in comparison with the MRTD with equally spaced event locations. The LSD was evaluated between the original LSF parameters, y(n), and the reconstructed LSF parameters, $\hat{y}(n)$.

TABLE 1. AVERAGE SPECTRAL DISTORTION OF INTERPOLATION METHODS

Event Locations	Avg. LSD
Tr, Tl, Nu	1.884
3 equally spaced events	1.918
Tr, Tl, Nu_1, Nu_2, Nu_3	1.712
5 equally spaced events	1.736

TABLE II: AVERAGE OF MEAN AND DEVIATION OF THE VALUE S

	LSF	Bark LSF	PLP-LSF	RASTA-LSF
Mean	170.46	123.25	115.37	140.84
Std	53.46	51.27	42.82	49.37

TABLE III: AVERAGE OF MEAN, DEVIATION AND MINIMUM OF MAHALANOBIS

	LSF	Bark LSF	PLP-LSF	RASTA-LSF
Mean	1.20	1.78	2.31	1.82
Std	0.31	0.46	0.62	0.73

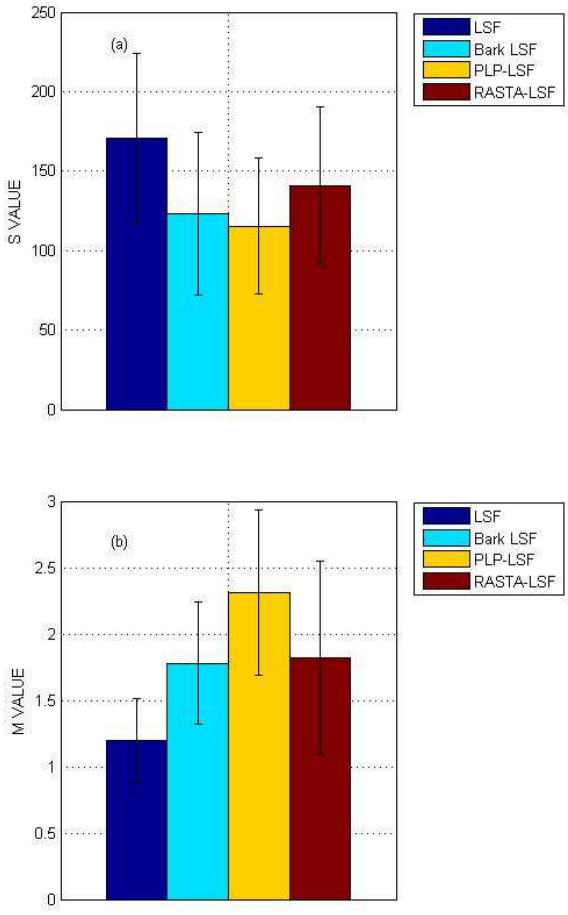

Fig.4. Stability of LSF targets (a) S Value (b) M Value

Table 1 gives the summary of spectral distortion results obtained. The results show that with the equivalent number of events, the interpolation performance of proposed method outperforms the method using equally events. This result supports that the even locations, estimated by local minima of STM and maxima of FSTM, were close with the static articulatory target locations, and could improve the interpolation performance.

C. Evaluating Stability of Spectral Targets by Intra-Category Target Variation

In order to quantitatively compare the stabilities of different kinds of spectral targets, mean and standard deviations of the value S, the sum of eigen values of the spectral targets covariance matrix were calculated [12, 13]. In our dataset, each single vowel has some variants, extracted from different CVC syllables.

Assuming that Γ_k is a covariance matrix of the targets sequences of the vowel *k*,

$$\Gamma_k = \sum_{n=1}^{N} (y_{nk} - \overline{y_k})(y_{nk} - \overline{y_k})^T \qquad (8)$$

Where N is number of vowel variants of the vowel k. $\overline{y}_k$ is the mean spectral targets of the vowel *k*. The value S_k is the sum of the eigen values ξ of the matrix Γ_k, that is,

$$S_k = \sum_i \xi_{ki} \quad (9)$$

If the variance of each vowel *k* decreases, the eigen values of matrix Γ_k decrease, and the value *S* also decreases. In our experiments, spectral targets were computed for only static nuclei intervals of phonemes. In order to compare the intra-category spectrum variation of the original LSF spectral target and PLP-LSFs spectral targets, mean and standard deviations of the value *S* in each category were calculated.

Table 1 and Fig. 4a shows the mean and standard deviation of the average value *S* for 13 Vietnamese single vowels individual categories, extracted from our Vietnamese dataset. This result indicates that the value *S* of the PLP-LSF is smallest, the original LSF is largest. Therefore, PLP-LSF was most robust under effects of neutralization, in comparison with original LSF and other variants.

D. Evaluating Stability of Spectral Targets by Inter-Category Target Variation

In order to evaluate the inter-category spectrum variation, an approximation of the Mahalanobis distance *M* was defined as follows [12, 13],

$$M = \frac{|b_i - b_j|}{\sigma_i + \sigma_j} \quad (10)$$

where b_i and b_j were the mean spectral targets of category *i* and *j* , and σ_i and σ_j were the standard deviations. The mean, standard deviation and minimum of the value *M* are shown in table 2, and Fig.4b, in which the value *M* with PLP-LSF is greatest and that with original LSF is smallest. Therefore, it is again confirmed that PLP-LSF is most robust under effects of neutralization, in comparison with original LSF and other variants.

VII. CONCLUSION

In this research, we used the STM, the FSTM, and the MRTD to model the coarticulation between intra-targets within nuclei intervals of phonemes, as well coarticulation between inter-targets of neighboring phonemes. The experimental results show that the proposed coarticulation model decomposed speech into context-insensitive spectral event targets, which are close with articulatory targets, and the context-sensitive spectral event functions, which closely represent the movements between the adjacent targets.

We also investigated the stability of spectral targets of vowel nucleus under effects of the vowel neutralization phenomenon in many different phonetic contexts. The experimental results show that the PLP-LSF target is a robust context-insensitive spectral target under effects of neutralization in comparison with original LSF as well as other combinations of PLP and LSF.

REFERENCES

[1] P. Delattre, "Coarticulation And The Locus Theory," *Studia Linguistica*, Vol. 23, Issue. 1, pp. 1–26, June 1969.

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

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

[4] Browman C.P, and Goldstein L, "Articulatory phonology: an with effects ofview," *Phonetica*, 49 (3-4), pp. 155-180, 1992.

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

[6] P.C. Nguyen, T. Ochi, M. Akagi, "Modified Restricted Temporal Decomposition and Its Application to Low Rate Speech Coding," *IEICE Trans. Inf. & 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] B.P. Nguyen, M. Akagi, "Phoneme-based Spectral Voice Conversion Using Temporal Decomposition and Gaussian Mixture Model", Proc. ICCE08, 2008.

[9] Hemamsky H, "Perceptual Linear Predictive (PLP) analysis of speech," *J Acoust Soc Am.* 87(4), pp. 1738 – 1752, 1990.

[10] Hermansky H, Morgan N, Bayya A, Kohn P, "RASTA-PLP speech analysis technique," *ICASSP 1992*, pp. 121 – 124, 1992.

[11] L.C. Mai, D.N. Duc, "Design of Vietnamese speech corpus and current status", *ISCSLP-06*, pp. 748-758, 2006.

[12] M. Akagi, "Evaluation of a Spectrum Target Prediction Model in Speech Perception," *J. of Acoust. Soc. Am*, Vol. 81, 1987.

[13] M. Akagi and Y. Tohkura, "Spectrum target prediction model and its application to speech recognition," *Computer Speech and Language 4*, pp. 325-344, 1990.