INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

The Parameterized Phoneme Identity Feature as a Continuous Real-Valued Vector for Neural Network based Speech Synthesis

Zhengqi Wen [1], Ya Li [1], Jianhua Tao [1,2]

[1] National Laboratory of Pattern Recognition,
[2]CAS Center for Excellence in Brain Science and Intelligence Technology,
Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China
{zqwen,yli,jhtao}@nlpr.ia.ac.cn

Abstract

In the speech synthesis systems, the phoneme identity feature indicated as the pronunciation unit is influenced by external contexts like the neighboring words and phonemes. This paper proposes to encode such relatedness and parameterize the pronunciation of the phoneme identity feature as a continuous real-valued vector. The vector, composed by a phoneme embedded vector (PEV) and a word embedded vector (WEV), is applied to substitute the binary vector whose representation is one-hot. It is realized in the word embedding model with the joint training structure where the PEV and WEV are learned together. The effectiveness of the proposed technique was evaluated by comparing it with the binary vector in the bidirectional long short term memory recurrent neural network (BLSTM-RNN) based speech synthesis systems. Improvement on the quality of the synthesized speech has been achieved from the proposed system, which proves the effectiveness of replacing the binary vector with the continuous real-valued vector in describing the phoneme identity feature.

Index Terms: phoneme embedded vector, word embedding, speech synthesis, BLSTM-RNN

1. Introduction

Statistical parametric speech synthesis uses statistical models to learn the mapping function from the contextual features to the speech parameters [1]. The hidden Markov model (HMM) [2] and the neural network (NN) [3] are two of the most popular statistical models in speech synthesis systems. In the HMM-based speech synthesis system [2], the speech parameters and duration of each HMM state are decided by the decision trees which are trained from the contextual features [4]. In the NN-based speech synthesis system [3], such as deep neural network (DNN) [5] and bidirectional long short term memory recurrent neural network (BLSTM-RNN) [6][7][8], the mapping function is trained directly from the contextual features to the speech parameters without the decision trees.

In these two synthesis methods, the contextual features include the phonetic information (phoneme identity feature) and the rest auxiliary information (part of speech, positional and prosodic features). Most of these contextual features are binary features suitable to construct the decision trees, whereas inappropriate to represent the input of the neural network. For phoneme identity feature, the consequence is that, the relatedness between two phonemes wouldn't be conveyed effectively supposing the phoneme identity feature is encoded as a binary vector with the one-hot representation in the input of the neural network. It is much more critical in natural language processing (NLP) where the dimension for one word is about tens of thousands. This problem has been alleviated with the putting forward of the word embedding model [9][10][11][12][13]. This model encodes a word as a real-valued low-dimensional vector based on the assumption that the semantic meanings of a word can be predicted from the external contexts with large-scale corpora. This assumption also works for the phonemes whose pronunciation is influenced by the neighboring words and phonemes. So this paper aims to study a continuous real-valued vector to parameterize the pronunciation of the phonemes. This vector will be used to substitute the binary vector of the phoneme identity feature in the input of the NN based speech synthesis systems.

The proposed vector is different from the vector space model (VSM) described in [14][15] which is trained from the matrix of co-occurrence statistics and further decomposed by singular value. Our vector consists of the phoneme embedded vector (PEV) and word embedded vector (WEV) which encodes the phoneme and word as a continuous real-valued vector. These two types of embedded vector are learned under the joint training structure in the word embedding model [16][17]. To testify the effectiveness of the proposed technique, we replace the binary vector with the proposed vector in the input of the BLSTM-RNN based speech synthesis systems to compare the performance of these two classes of vectors. The objective measure shows the root mean square error (RMSE) of the generated line spectral pair (LSP) is reduced by about 5% and the subjective listening test also indicates that the synthesized speech is preferred by about 30% for the continuous real-valued vector than the binary vector. Thus, the results demonstrate the outstanding representation ability of the proposed vector in describing the phoneme identity feature in the NN based speech synthesis.

2. Phoneme Identity Feature Parameterization

The word embedding is proposed to encode the semantic meaning of a word as a continuous real-valued vector in the NN based language model, while one of the problems is that the training process is too time consuming. So recently NLP researchers proposed a number of the word embedding models to train the word embedding directly, such as Global C&W [18], continuous bag-of-words model (CBOW) [13] and Skip-Gram [13]. The following will take CBOW as an example to jointly train the phoneme embedded vector (PEV) and the word embedded vector (WEV).

http://dx.doi.org/10.21437/Interspeech.2016-222

2.1. CBOW

CBOW is a simple neural network described in Fig. 1. This network predicts the target word directly by the neighboring context words from a sliding window. The structure of this network keeps only an input layer and an output layer without the hidden layers and activation function. Therefore, the training of the CBOW is very efficient.

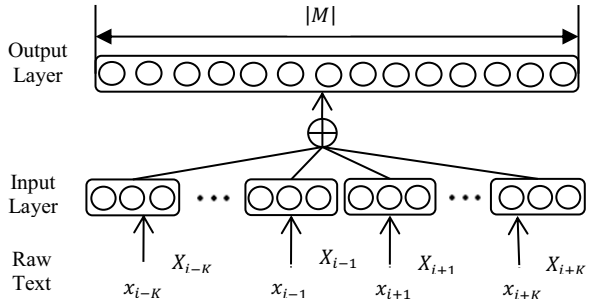

Figure 1: *The block diagram of CBOW.*

Given a sentence with N training words $S = \{x_1, x_2, \cdots, x_N\}$, the object of training the CBOW is to maximize the average log probability in Equation (1).

$$\mathcal{L}(S) = \frac{1}{N-2K}\sum_{i=K+1}^{N-K} logP(x_i|x_{i-K}, \cdots, x_{i+K}) \quad (1)$$

where K is the size of the sliding window for the neighboring context words. The formulation for the probability $P(x_i|x_{i-K}, \cdots, x_{i+K})$ is a softmax function described in Equation (2).

$$P(x_i|x_{i-K}, \cdots, x_{i+K}) = \frac{\exp(X_0^T \cdot X_i)}{\sum_{x_j \in W} \exp(X_0^T \cdot X_j)} \quad (2)$$

where W is the word vocabulary, X_i is the WEV of the target word x_i, and X_0 is the average of all neighboring context words in Equation (3).

$$X_0 = \frac{1}{2K}\sum_{j=i-K,\cdots,i+K, j\neq i} X_j \quad (3)$$

The gradient descent algorithm is used to learn the parameters of the neural network. From Equation (1-3), the only unknown parameter is the WEV which can be updated according to the gradient of the objective function in Equation (1) with a learning rate λ in Equation (4).

$$X_i = X_i + \lambda \cdot \frac{\partial \mathcal{L}(S)}{\partial X_i} \quad (4)$$

2.2. Joint Training for PEV and WEV

The word embedding is learned based on the assumption that the semantic meaning of a word is influenced by the neighboring contextual information and can be predicted by the neighboring context words under a large corpus. Similarly, according to the phonetics [19], the phonemes, constituting the pronunciation of the word, are also influenced by the neighboring contextual information, such as words and phonemes. So we can believe that encoding the pronunciation of the phoneme as a continuous real-valued vector is accessible and impactful.

In English, every word is composed of the phonemes. Differently, Mandarin is a monosyllabic language, where each character constitutes a single syllable. Each syllable consists of an optional initial (consonant), a final (vowel) and a lexical tone. So the trained unit for the pronunciation of the word is phoneme for English and initial or final for Mandarin. They are both named as phoneme embedded vector (PEV) in this paper.

It is difficult to train the PEV directly from the large corpus because the PEV takes a non-semantic meaning of a word. But it could be learned simultaneously with the WEV in a joint training structure described in [16][17]. This structure is proposed to train the syllable embedded vector with the WEV for Mandarin where the target word is predicted by the context word and the composition syllables together. This paper replaces the composition syllables by the pronunciation initial and final, and predicts the target word by the context word and the pronunciation initial and final together. In this way, the PEV can be learned together with WEV. So the embedded vector for the context word x_i is changed from X_i to X_i^{new} in Equation (5). The update for the PEV is kept the same as WEV in Equation (4).

$$X_i^{new} = X_i + \frac{1}{N_i}\sum_{m=1}^{N_i} P_m \quad (5)$$

where X_i^{new} is the composed embedded vector, X_i is the word embedded vector (WEV), P_m is the phoneme embedded vector (PEV), N_i is the number of initial and final for ith word.

3. Integration in the BLSTM-RNN based speech synthesis system

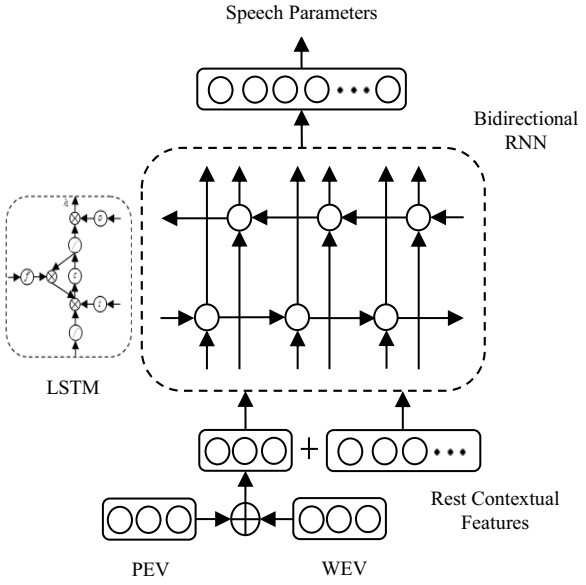

Figure 2: *The framework of the bidirectional long short term memory recurrent neural network (BLSTM-RNN) based speech synthesis with the input of the PEV and WEV.*

Word embedding has been proved to be helpful in describing the contextual features for the speech synthesis systems in [20]. It is used as an additional contextual feature for the BLSTM-RNN based speech synthesis system and the synthesized speech quality has been improved with the additional feature. This paper also adopts the word embedding in the speech synthesis systems, whereas the word embedding is utilized to describe the pronunciation of the phonemes with the PEV trained in Section 2.

In the HMM and NN based speech synthesis systems, resulting from the influence of the surrounding context, the pronunciation of the word is represented as a number of phoneme identity features. The number of the phoneme

identity features is usually set as 5. This representation is suitable for the HMM-based speech synthesis system where the decision trees use the binary input. But encoding the phoneme identity features as binary vectors with one-hot representation is not appropriate for the input of the NN-based speech synthesis system where the relatedness between the surrounding phonemes is not considered. The proposed PEV trained from the neighboring contexts takes the relatedness into consideration and the representation as continuous real-valued vectors is also suitable for the input of the NN based speech synthesis systems.

The proposed BLSTM-RNN based speech synthesis system which integrated with the PEV and WEV for the pronunciation of the phonemes is described in Fig. 2. The binary vector of the phoneme identity feature in the contextual feature is replaced by a combination of the PEV and WEV described in Equation (6).

$$\mathrm{PI}_{i,m} = \mathrm{X}_{\mathrm{i}} + \mathrm{P}_{i,m} \quad (6)$$

where $\mathrm{PI}_{i,m}$ is the continuous real-valued vector for the phoneme identity feature, X_{i} is the WEV of ith word, $\mathrm{P}_{i,m}$ is the PEV of the m the phoneme in the ith word.

The remaining input contextual features include a binary vector for the part of speech (POS), a binary vector for the lexical tone and the numerical features for the numerical contexts, e.g., the number of syllables in the word or the position of the current frame in the current phoneme.

The mapping function from the input contextual features to the output features is trained sequence by sequence with the generative characteristic of the BLSTM-RNN. The output features include the logarithm of the fundamental frequency (LF0), voiced/unvoiced flag and line spectral pairs (LSP) [21] with first and second order difference. In the synthesis stage, the parameter generation algorithm [22] and the vocoding technique [23] were used to synthesis the speech.

4. Experiments and Discussion

To evaluate the proposed technique, a series of experiments have been carried out for the Mandarin language. Firstly, the word embedding after joint training with the phoneme embedded vector is compared with other training strategy. Then, the replacement from the binary vector with the one-hot representation to the PEV and WEV for the phoneme identity features was made for evaluation in the BLSTM-RNN based speech synthesis systems. After that, the PEV trained with the lexical tone for the tone language of Mandarin was compared by the generated voice with the PEV without the lexical tone.

The corpus used in the following experiments is about seven hours from a female talker speaking mandarin. The speech parameters include the LSP extracted from the STRAIGHT spectrum [24] and the logarithm of the fundamental frequency (LF0). The contextual features are represented as a vector for every frame with a dimension of 379 where 300 dimensions are used for the one-hot representation of the phoneme identity features. In the proposed BLSTM-RNN based speech synthesis system, the phoneme identity feature is represented with the continuous vector generated from the PEV with WEV in Equation (6). So the dimension for training the PEV with the WEV is also set to 300 to keep the same dimension with the input contextual feature. Before taken into the network training, these two features are both normalized as a Gaussian distribution with zero mean and unity variance. The quality of the synthesized speech was verified in two ways. The first one is the objective measures, including the root mean square error (RMSE) between the generated and the original speech parameters and log spectral distance (LSD) [25] between the generated and the original waveforms. The other is a subjective measure in terms of the ABX preference scores [26] where subjects were asked to listen to two versions of synthesized speech and choose the preferred one according to quality and naturalness of the synthesized speech. The better one will get a preference score of "1" or no preference (N/P) score of "1". The final scores are calculated by the mean value of the scores given by the 20 listeners who are working in the speech technology areas.

4.1. Word Embedding with Different Training Strategy

We collected a large Mandarin corpus for embedding training. The corpus contains 240 million words and the vocabulary size is 238 thousand. The training tool provided in [17] is a joint training version of wrod2vec and is further adopted to train the embedding vectors in this paper. Three versions of word embedding have been trained for evaluation: word embedded vector (WEV) only, word embedded vector (WEV) with syllable embedded vector (SEV) and word embedded vector (WEV) with phoneme embedded vector (PEV). In Table 1, the words with the five lowest distances from the word of "Beijing" are listed for these three versions of word embedding. The words listed are all the provincial capitals of China where the top two cities of "Tianjing" and "Shanghai" are both the municipalities as is "Beijing" in China. The results indicate that the word embedding has been learned successfully from the joint training structure with WEV and PEV.

Table 1. *The words with the five lowest distances from "Beijing" in three versions of word embedding.*

WEV	WEV with SEV	WEV with PEV
Tianjing: 0.790455	Shanghai: 0.768427	Tianjing: 0.788772
Shanghai: 0.793893	Tianjing: 0.783911	Shanghai: 0.790337
Guangzhou: 0.832642	Chengdu: 0.810766	Chengdu: 0.828384
Chengdu: 0.836554	Nanjing: 0.816841	Nanjing: 0.833955
Nanjing: 0.837879	Hangzhou: 0.822146	Guangzhou: 0.838408

The PEV takes a non-semantic meaning of the word and direct comparison between the PEV for different phonemes cannot get a meaningful result. But the PEV together with the WEV is assumed to be able to describe the pronunciation of the word because the pronunciation is also influenced by the neighboring contexts. This assumption will be testified in the BLSTM-RNN based speech synthesis systems with the input of PEV and WEV.

4.2. Replacement from one-hot representation to the PEV and WEV

Two versions of the BLSTM-RNN based speech synthesis systems were trained. One takes the binary vector with the one-hot representation for the phoneme identity features as the

input contextual feature. The other replaces the binary vector with the continuous real-valued vector composed by the PEV and WEV in the input contextual feature. After replacement, the dimension of the PEV and WEV is consistent with the binary vector with one-hot representation for the phoneme identity features. After training, two versions of synthesized speech were generated to be compared in the objective measures and the preference listening test. The experimental results are described in Table 2 and Table 3.

Three types of objective measures are considered in Table 2. All these measures show that the BLSTM-RNN based speech synthesis system with the input of the continuous real-valued vector generates a better trajectory of the speech parameters than the binary vector. The biggest improvement is the RMSE of LSP which has been reduced by about 5% from 1.113 to 1.053. The other two measures are also showing a small reduction.

The generated trajectory of the speech parameters was further compared in the subjective listening test. The ABX preference score is listed in Table 3. About 39.3% of listeners preferred the speech synthesized with the input of the continuous real-valued vector whereas about 25.3% of listeners chose the better speech synthesized with input of the binary vector. The rest, about 35.4, were without preference for either version of the synthesized speech.

It can be concluded from the results of the objective measures and the subjective listening test, that the continuous real-valued vector composed by the PEV and WEV is a better substitution compared to the binary vector of the phoneme identity feature in the input contextual feature.

Table 2. *The LSD, RMSE for LF0 and RMSE for LSP for the comparison between the binary vector with one-hot representation and the continuous real-valued vector in the BLSTM-RNN based speech synthesis systems.*

	LSD	LF0	LSP
One-Hot	4.899	0.133	1.113
PEV&WEV	4.885	0.132	1.053

Table 3. *Preference scores with confidence interval for the comparison between the binary vector with the one-hot representation and the continuous real-valued vector in the BLSTM-RNN based speech synthesis systems.*

One-Hot	PEV&WEV	N/P
0.253 ($\pm$0.13)	0.393($\pm$0.09)	0.354($\pm$0.11)

4.3. Comparison between PEV and PEV with the tone

Mandarin is a tone language and there is a binary vector with a dimension of 5 to distinguish different tones of the Mandarin syllable in the input contextual feature. In the previous experiments, the phoneme embedded vector (PEV) was trained without the tone tag. In the following experiments, the PEV will be trained with the tone tag named as Tone-PEV.

In the previous experiment, the number of the PEV is 60 and the number of the initial and final in Mandarin is also 60. In training the Tone-PEV, the number will be expanded to 300 with the tone tag while the dimension for the Tone-PEV is kept as the same. The experiment was carried out to compare these two versions of the PEV in the BLSTM-RNN based speech synthesis systems. The results of the comparison are described in Table 5 and Table 6.

The objective measures demonstrate that the Tone-PEV is not as good as the PEV in the BLSTM-RNN based speech synthesis systems, especially for the RMSE of LF0 which has been increased from 0.132 to 0.164. The subjective listening tests also indicate that the generated voice for the PEV without the tone is more favored by about 60% than with the Tone-PEV. therefore, we can draw the conclusion that it is not appropriate to describe the tone information in the embedded vector. One possible reason for this is that the proposed real-valued vector is a combination with the PEV and WEV. But the WEV cannot be learned with one tone tag. So only considering the PEV with the tone tag is not enough to describe the tone information.

Table 4. *The LSD, RMSE for LF0 and RMSE for LSP for the phoneme embedded vector and the phoneme embedded vector with the tone tag in the BLSTM-RNN based speech synthesis systems.*

	LSD	LF0	LSP
PEV	4.885	0.132	1.053
Tone-PEV	5.234	0.164	1.113

Table 5. *Preference scores with confidence interval between the phoneme embedded vector and the phoneme embedded vector with the tone tag in the BLSTM-RNN based speech synthesis systems.*

PEV	Tone-PEV	N/P
0.743($\pm$0.12)	0.144($\pm$0.08)	0.113($\pm$0.15)

5. Conclusion and Future Work

This paper proposes a method to encode the pronunciation of the phoneme identity feature as a continuous real-valued vector. It is realized by the combination of the phoneme embedded vector (PEV) and word embedded vector (WEV) which learned under the joint training structure in the word embedding model. This vector is utilized to replace the binary vector of the phoneme identity feature in the input of the bidirectional long short term memory recurrent neural network (BLSTM-RNN) based speech synthesis systems. Experiments were carried out to testify the effectiveness of the proposed technique and the results show that the quality of the synthesized speech has been improved by replacing the binary vector with the continuous real-valued vector. It is proved that the continuous real-valued vector is able to parameterize the pronunciation of the phoneme.

6. Acknowledgements

This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No.2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61425017, No.61403386), the Strategic Priority Research Program of the CAS (GrantXDB02080006) and the Major Program for the National Social Science Fund of China(13&ZD189).

7. References

[1] H. Zen, K. Tokuda and A.W. Black, “Statistical parametric speech synthesis”, Speech Communication, 51(11):1039-1064, 2009.

[2] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black and K. Tokuda, “The HMM-based speech synthesis system version 2.0”, in Proc. of ISCA SSW6, 2007.

[3] O. Karaali, G. Corrigan and I. Gerson, “Speech synthesis with neural networks”, in Proc. World Congress on Neural Networks, pp. 45-50, 1996.

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis”, in Proc. of Eurospeech, pp.2347-2350, 1999.

[5] H. Zen, A. Senior and M. Senior, “Statistical Parametric Speech Synthesis Using Deep Neural Networks”, In Proc. ICASSP, pp.8012-8016, 2013.

[6] Y. Fan, Y. Qian, F. Xie, and F. K.Soong, “TTS synthesis with bidirectional LSTM based recurrent neural networks”, in Proc. of INTERSPEECH, 2014.

[7] S. Mike and K. Paliwal. “Bidirectional recurrent neural networks”, IEEE Transactions on Signal Processing, 45(11): 2673-2681, 1997.

[8] H. Sepp and S. Jürgen, “Long short-term memory”, Neural Computation, 9(8):1735-1780, 1997.

[9] Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin, “A Neural probabilistic language model”, Journal of Machine Learning Research, pp. 1137-1155, 2003.

[10] T. Mikolov, M. Karafiat, L. Burget, J. “Honza” Cernocky and S. Khudanpur, “Recurrent neural network based language model”, in Proc. of INTERSPEECH, pp. 1045-1048, 2010.

[11] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, “Natural language processing (almost) from scratch”, Journal of Machine Learning Research, 12:2493–2537, 2011.

[12] E. H. Huang, R. Socher, C. D. M. and A. Y. Ng, “Improving word representations via global context and multiple word prototypes” , in Proc. of ACL, pp. 873–882 ,2012.

[13] T. Mikolov, K. Chen, G. Corrado and J. Dean, “Efficient estimation of word representations in vector space”, in Proc. of CoRR, 2013.

[14] H. Lu, S. King, and O. Watts, “Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis”, in the 8th ISCA Speech Synthesis Workshop, pp. 281-285, 2013.

[15] O. Watts, “Unsupervised learning for text-to-speech synthesis”, Ph.d. dissertation, 2013.

[16] X. Chen, L. Xu, Z. Liu, M. Sun, and H, Luan."Joint learning of character and word embeddings", in International Joint Conference on Artificial Intelligence, 2015

[17] F, Sun, J. Guo, Y. Lan, J. Xu, and X. Cheng, "Inside out: Two jointly predictive models for word representations and phrase representations", in Proceedings of the 30th AAAI conference, 2016.

[18] R. Collobert, and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning", in International Conference on Machine Learning, 2008.

[19] Z. J. Wu, "The Chinese Phonetics in ‘Man-Machine Dialogue’", Chinese Teaching In The World, vol4, pp3-20,1997 (In Chinese).

[20] P. Wang, Y. Qian, F. K. Soong, L. He and H. Zhao, “Word Embedded for Recurrent Neural Network based TTS Synthesis”, in Proc. of ICASSP, pp. 4879-4883, 2015.

[21] F. K., Soong, and B. H., Juang, “Line spectrum pair (UP) and speech data compression”, in Proc. of ICASSP, pp. 1.10.1-1.10.4, 1984.

[22] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis”, in Proc. of ICASSP, pp.1315-1318, June 2000.

[23] Z. Wen, J. Tao, S. Pan and Y. Wang, “Pitch-Scaled Spectrum Based Excitation Model for HMM-based Speech Synthesis”, Journal of Snigal Processing Systems, 74(3) :423-435, 2014.

[24] H., Kawahara, I., Masuda-Katsuse amd A. de Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds”, Speech Communication, 27(5), 187–207, 1999.

[25] Y. Ephraim and D. Malah, “Speech Enhancement using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator”, IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSp-33, No. 2, pp. 443-445, 1985.

[26] L. Blin, O. Boeffard and V. Barreaud, “WEB-based listening test system for speech synthesis and speech conversion evaluation”, in Proc. of LREC (Marrakech (Morocco)), 2008.