# MAXIMUM INTELLIGIBILITY-BASED CLOSE-LOOP SPEECH SYNTHESIS FRAMEWORK FOR NOISY ENVIRONMENTS

*Yuan-Fu Liao, Ming-Long Wu and Jia-Chi Lin*

Department of Electronic Engineering, National Taipei University of Technology, No. 1, Section 3,
Chung-hsiao East Road, Taipei 10608, Taiwan
yfliao@ntut.edu.tw, a222604418@yahoo.com.tw, kiki90235@hotmail.com

## ABSTRACT

This paper proposes a maximum intelligibility (MI)-based close-loop speech synthesis framework to actively compensate for the distortion of background noises. In this framework, an extra environmental noise-sensing microphone and an automatic speech recognition (ASR) module are utilized to approximate a subjective intelligibility measure. The hidden Markov model-based speech synthesis system (HTS) is then online adjusted by using the MI-based model adaptation algorithm. Experimental results of two subjective listening tests in noisy environments show that the proposed approach obtains 64% of the votes in an A/B preference test and helps the participants reduce word dictation errors by relative 26% when compared to an HTS baseline.

***Index Terms***— Speech synthesis, speech intelligibility, automatic speech recognition, minimum classification error

## 1. INTRODUCTION

Traditional text-to-speech (TTS) systems usually focus more on improving the naturalness and similarity of synthesized speech and less on intelligibility in noisy environments. However, in real-life TTS applications, background noise is often unavoidable. For example, in scenarios involving GPS car navigation or mobile phone screen reader for visually impaired people, there is often strong environmental noise. In particular, the output volume of a TTS system may be limited or even less than the background noise in these situations.

To alleviate this problem, many approaches, such as spectral tilt [1-3], format enhancement [4], waveform companding [5], speech transformation [6] and source-filter model modification [7] are quite helpful. However, such methods are often open-loop or post-processing procedures. In other words, conventional TTS systems are deaf talkers because they are not aware of environmental noise.

By contrast, humans are listening talkers and have the capacity to perceive background noise and adjust their voices to transmit messages efficiently. For example, the Lombard effect [8] is the involuntary tendency of speakers to increase their vocal efforts when speaking in loud noise to enhance the audibility of their voices.

Therefore, this paper proposes a maximum intelligibility (MI)-based close-loop speech synthesis framework to mimic human ability and develop the capacity of a TTS system to actively compensate for the distortion of background noise. In brief, the proposed speech synthesis framework is equipped with an extra noise-sensing microphone and a speech recognizer to feedback speech intelligibility measure and is, therefore, a close-loop synthesis-by-analysis approach. And it should be possible to apply the proposed framework to simultaneously adjust the spectral, pitch and duration of synthesized speech [6-7]. However, in this paper, the proposed framework will be implemented to only modify the spectral parameters of HTS [9] phone models.

## 2. MAXIMUM INTELLIGILITY SPECCH SYNTHESIS FRAMEWORK

Fig. 1 shows the design of the proposed MI-based framework. The major differences from conventional TTS are (1) an extra environmental noise-sensing microphone; (2) an ASR module; (3) a subjective intelligibility measure; and (4) an MI-based online model adaptation algorithm. These modules are combined to simulate the functions of the human ear and brain.

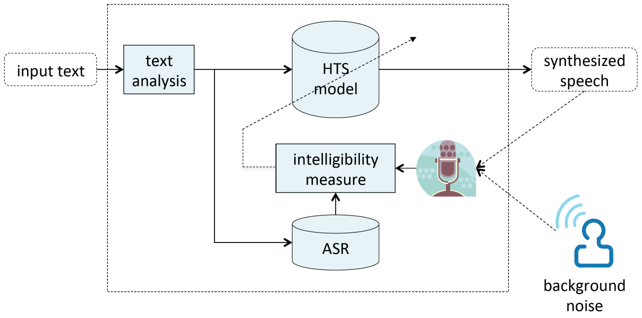


Fig. 1: Block diagram of the proposed maximum intelligibility-based close-loop speech synthesis framework for noisy environment.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

The operation of this framework is is run in utterance-by-utterance mode and could be described as follows: (1) background noise are online picked up by the microphone and mixed with a synthesized speech utterance; (2) the ASR module aligns (since we knew the input text) and recognizes the mixed noisy speech to determine a correct and a most competitive hypotheses, respectively; (3) a smooth recognition error rate function is computed based on both hypotheses to approximate the subjective intelligibility measure; (4) the HTS phone models used to synthesize the utterance are then adjusted according to the guide of the measurement, and the utterance is re-synthesized; finally (5) the entire procedure is iterated until observed speech intelligibility measure could not be further improved.

It is worth noting that in this operation, only a subset of HTS phone models selected to synthesize the input sentence is involved in the adaptation procedure. Therefore, the additional computation time required is somehow acceptable. And the ASR module could be developed using (1) a large multi-speaker corpus for training a speaker-independent acoustic model, or (2) the same single-speaker corpus for training the HTS voice. These two cases could be treated as two types of subjective intelligibility feedback provided by other people or the speaker himself.

## 3. MAXIMUM INTELLIGILITY ALGORITHM

To simplify the computation complexity and directly adjust the spectral parameters of HTS phone models, the proposed algorithm is implemented in the mel-frequency cepstral coefficient (MFCC) domain, as illustrated in Fig. 2.

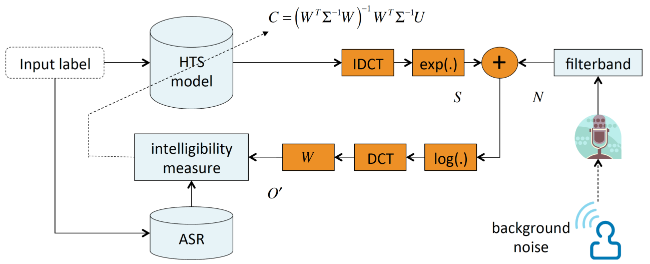


Fig. 2: Block diagram of the implementing the proposed MI-based approach to directly adjust the spectral parameters of HTS phone models.

In the synthesis phase, the most suitable HTS phone model sequence for synthesizing an input sentence are chosen and concatenated to form a mixture mean vector sequence $U=[\mu_{q_1}^{HTS},...,\mu_{q_t}^{HTS},...,\mu_{q_T}^{HTS}]$. The mean vector sequence is then transformed into an optimal output MFCC vector sequence $C=[c_1,...,c_t,...,c_T]$ using the parameter generation algorithm [10] as Eq. (1).

$$C=\left(W^T\Sigma^{-1}W\right)^{-1}W^T\Sigma^{-1}U \tag{1}$$

Here $W$ and $\Sigma$ are the dynamic feature generation window function and the covariance matrices of HTS phone models.

In the analysis phase, the output MFCC vector sequence $C$ is converted into a linear spectral vector sequence $S$ using the inverse cosine transform (IDCT) and exponential function, i.e., $S=\exp\left(IDCT\left(C\right)\right)$. Then, the spectrum of observed background noise $N$ is online mixed with $S$ and transformed by using a logarithmic function, a discrete cosine transform (DCT) and the window function $W$ into a noisy feature vector sequence $O'=[o'_1,...,o'_t,...,o'_T]$ as Eq. (2) to reflect the distortion of environmental noise.

$$O'=W\left(DCT\left(\log\left(S+N\right)\right)\right) \tag{2}$$

Then the ASR (with an acoustic model $\Lambda^{ASR}$) module aligns and recognizes the noisy MFCCs to generate a correct and a most competitive phone model state sequences $q$ and $q^*$ with corresponding likelihood scores $g\left(\cdot\right)$ and $g^*\left(\cdot\right)$ as Eq. (3) and (4) respectively.

$$\begin{aligned} g\left(O',q;\Lambda^{ASR}\right) &= \log P\left(O',q\,|\,\Lambda^{ASR}\right) \\ &\approx -\sum_{t=1}^{T}\sum_{j=1}^{J}\frac{\left(o'_{t,j}-\mu_{q_t,j}^{ASR}\right)^2}{2\sigma_{q_t,j}^{ASR}} \end{aligned} \tag{3}$$

$$\begin{aligned} g^*\left(O',q^*;\Lambda^{ASR}\right) &= \log P\left(O',q^*\,|\,\Lambda^{ASR}\right) \\ &\approx -\sum_{t=1}^{T}\sum_{j=1}^{J}\frac{\left(o'_{t,j}-\mu_{q_t^*,j}^{ASR}\right)^2}{2\sigma_{q_t^*,j}^{ASR}} \end{aligned} \tag{4}$$

Here we assume that a partitioned Gaussian density function is adopted, $J$ is the size of mean vector and $\sigma_{q_t,j}^{ASR}$ is the $q_t$-th state, $j$–th dimension variance of the Gaussian mixture.

A misclassification function $D\left(\cdot\right)$ is then defined using Eq. (5) and further transformed by using a zero-one sigmod function $L\left(\cdot\right)$ (with parameters $\alpha$ and $\beta$) to approximate a smooth recognition error rate, as described in Eq. (6).

$$D=-g\left(O',q;\Lambda^{ASR}\right)+g^*\left(O',q^*;\Lambda^{ASR}\right) \tag{5}$$

$$L\left(D\right)=\frac{1}{1+\exp(-\alpha D+\beta)} \tag{6}$$

This smooth error rate function is finally used to approximate the subjective intelligibility measure.

Thus, the parameters of the HTS phone models will be iteratively adjusted using the probabilistic gradient decent (GPD) [11-12] method, as defined in Eq. (7) (only the mean vectors are considered here).

$$\mu_{i,j}^{HTS}(n+1) = \mu_{i,j}^{HTS}(n) - \varepsilon\left(1 - \frac{n}{n_{\max}}\right)\frac{\partial L(D)}{\partial \mu_{i,j}^{HTS}} \quad (7)$$

Here, $n$ is the iteration index, $\mu_{i,j}^{HTS}$ is the $i$-th state, $j$–th dimension Gaussian mean, $\varepsilon$ is the learning step, and $n_{\max}$ is the maximum number of iterations.

In more detail, the chain-rule is applied to the last term of Eq. (7):

$$\frac{\partial L(D)}{\partial \mu_{i,j}^{HTS}} = L(D)(1 - L(D))\frac{\partial D}{\partial \mu_{i,j}^{HTS}} \quad (8)$$

$$\frac{\partial D}{\partial \mu_{i,j}^{HTS}} = -\frac{\partial g(O', q; \Lambda^{ASR})}{\partial \mu_{i,j}^{HTS}} + \frac{\partial g^*(O', q^*; \Lambda^{ASR})}{\partial \mu_{i,j}^{HTS}} \quad (9)$$

$$\frac{\partial g(O', q; \Lambda^{ASR})}{\partial \mu_{i,j}^{HTS}} = \sum_{t \& q_t = i} \frac{\left(o'_{t,j} - \mu_{i,j}^{ASR}\right)}{\sigma_{i,j}^{ASR}} \frac{\partial o'_{t,j}}{\partial \mu_{i,j}^{HTS}} \quad (10)$$

$$\frac{\partial o'_{t,j}}{\partial \mu_{i,j}^{HTS}} = \frac{\partial o'_{t,j}}{\partial c_{t,j}} \frac{\partial c_{t,j}}{\partial \mu_{i,j}^{HTS}} \quad (11)$$

Moreover, the term $\frac{\partial o'_{t,j}}{\partial c_{t,j}}$ in Eq. (11) will be approximated by using a first-order Taylor expansion formulation.

In summary, it could be found, especially from Eq. (10), that the mechanism of the proposed MI-based algorithm is to move the noisy MFCC, $o'_{t,j}$, toward the correct reference ASR model mean, $\mu^{ASR}{}_{i,j}$ and away the most competitive one. In other words, it alleviates the distortions of MFCCs in the presence of background noises (will be further discussed in next section). Therefore the intelligibility of synthesized speech could be increased in noisy environments.

## 4. EXPERIMENTAL RESULTS

The performance of the proposed approach was evaluated and compared with a HTS baseline using a single female speaker Mandarin Chinese corpus released in the Blizzard Challenge 2009 [13] and NoiseX-92 [14] database.

For all experiments, an HTS system similar to our NTUT entry [15] to the Blizzard Challenge 2009 was first established as the baseline system and then modified utterance-by-utterance by using the proposed MI-based approach. In all the following experiments, four different background noises, including babel, factory, tank and white, at 0, -5 and -10 dB segmental signal-to-noise ratio (SNR) conditions were tested. The ASR acoustic model used is same as the HTS one. The energies of all synthesized utterances were first normalized and then mixed with noises.

Two subjective tests were evaluated: (1) A/B preference test on 480 news utterances and (2) word dictation test on 480 semantically unpredictable sentences (SUS). Those utterances were selected from the Blizzard Challenge 2009 test set. All subjective tests involved 24 native speakers of Mandarin Chinese with random assignment (20 news pairs and 20 SUS utterances per participant).

For the A/B preference test, the participants were asked to compare two synthesized utterances generated by the HTS baseline and the proposed MI-based method (presented in random order) and choose the more intelligible one pair-by-pair. For the word diction test, each participant had to transcribe 20 SUS sentences utterance-by-utterance. All listening tests were done in a quiet office room and equipped with a high quality circumaural headphone.

First of all, Fig. 3 shows the spectrums of the two synthesized utterances of the same test sentence in the presence of strong (SNR=-10 dB) tank noise before and after applying the proposed MI-based algorithm, respectively. It could be found from this typical example that the proposed MI-based method automatically shifted most of the energy of the synthesized speech to higher frequency band to keep away from the lower frequency band tank engine noise. It also enhanced the formants of synthesized speech in higher frequency band. It therefore could mask the background noise better than the HTS baseline to alleviate the distortions of MFCCs (especially, the MFCC norm shrinking problem) in the presence of background noises. This may confirm the benefits in using the proposed MI-based method.

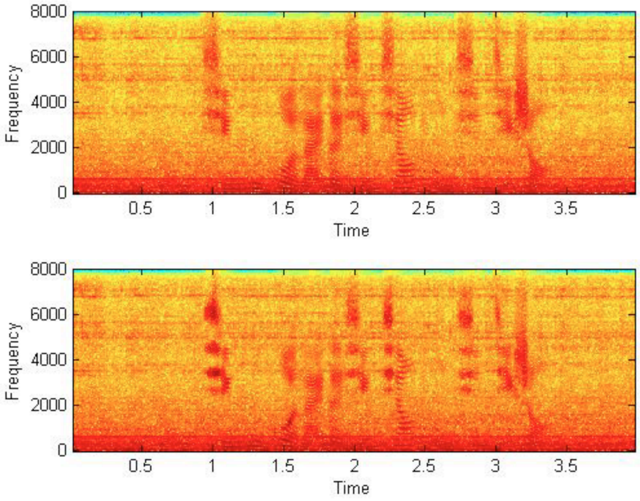
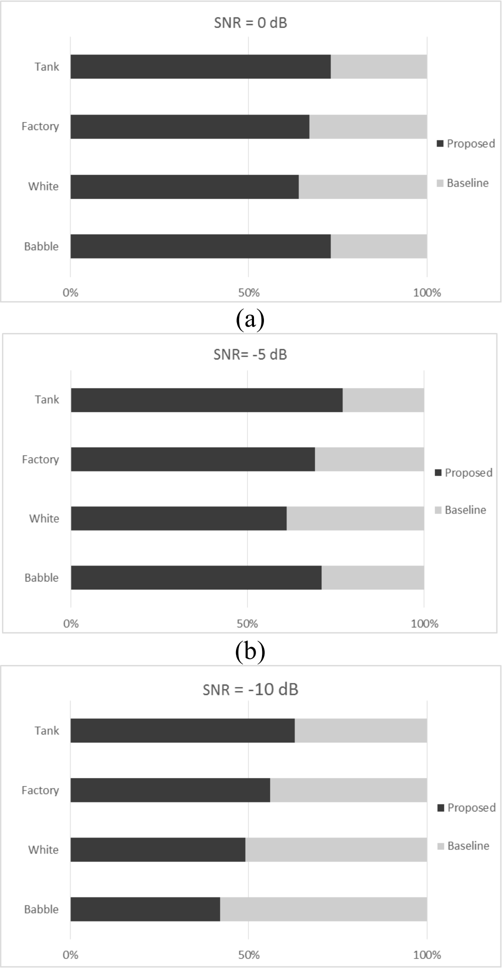


Fig. 3: The spectrograms of the two synthesized utterances of the same test sentence in the presence of strong (SNR=-10 dB) tank noise before (top panel) and after (bottom panel) applying the proposed MI-based algorithm (x- and y-axis are in second and Hz).

Secondary, Fig. 4 (a)~(c) show the results of the subjective A/B preference test on the 4 different background noises and three SNR (0, -5 and -10 dB) conditions. On average (over all SNRs and noise types), the proposed MI-based approach obtained 64% of the votes. In other words, the participants believed that the synthesized utterances generated by the MI-based system were more intelligible than the HTS baseline ones. The only two exceptions are the cases of strong (-10 dB) babble and white noise. This may due to the larger spectrum overlapping between speech and babble and white background noises.



(a)

(b)

(c)

Fig. 4: A/B preference test results of the proposed MI-based approach and the HTS baseline on four different background noises and three SNR conditions: average votes over (a) 0 dB, (b) -5 dB and (c) -10 dB cases.

Finally, Table 1 lists the results of the word dictation test on the 4 different background noises and three SNR (0, -5 and -10 dB) conditions, showing that the proposed MI-based approach helped participants to dictate more words correctly (word-error rate was reduced from 29.50% to 21.86%, which is about 26% relative error rate reduction). These results may further confirm the superiority of the proposed MI-based method.

Table 1: Word dictation test results of the proposed MI-based approach and the HTS baseline on four different background noises and three SNR conditions: average word-error rate (%) over 0, -5, and -10 dB SNR conditions.

| SNR=0 dB | HTS baseline | Proposed method |
|---|---|---|
| Babble | 18.18 | 9.09 |
| White | 27.27 | 20.45 |
| Factory | 20.45 | 13.63 |
| Tank | 15.90 | 11.36 |
| Average | 20.45 | 13.63 |
| SNR=-5 dB | HTS baseline | Proposed method |
| Babble | 27.27 | 22.27 |
| White | 34.09 | 25.00 |
| Factory | 27.27 | 18.18 |
| Tank | 27.27 | 15.90 |
| Average | 28.98 | 20.33 |
| SNR=-10 dB | HTS baseline | Proposed method |
| Babble | 40.90 | 34.09 |
| White | 52.27 | 45.45 |
| Factory | 36.36 | 27.27 |
| Tank | 31.81 | 25.00 |
| Average | 40.34 | 32.95 |

## 5. RELATION TO PRIOR WORK

The work presented here has focused on a close-loop synthesis-by-analysis algorithm that integrates both TTS and ASR modules. The works by other people usually consider only the TTS itself and are open-loop or post-processing approaches. For example, Valentini-Botinhao et al [2-3] apply Glimpse Proportion (GP) measure to modify the spectral envelope of synthesized speech during speech parameter generation phase. The work by Huang, et al [7] is based on speech-to-Lombard speech transformation in a post-processing stage. Suni et al [6] build a source-filter model (GlottHMM) and increase speaker's vocal efforts by empirically adjusting related parameters.

## 6. CONCLUSIONS

The experimental results of two subjective listening tests show that the proposed MI-based method improved the intelligibility of the synthesized speech in various noisy environments. Especially, it obtains 64% of the votes in an A/B preference test and helped the participants reduce word dictation errors by relative 26% error reduction when compared to the HTS baseline. In the future, the proposed framework will be extended to adjust the pitch and durations of synthesized speech.

## 7. ACKNOWLEDGMENTS

This work was supported by the National Science Council (NSC), Taiwan with contract 101-2221-E-027-129.

## 8. REFERENCES

1. Daniel Erro, Yannis Stylianou, Eva Navas and Inma Hernaez, “Implementation of Simple Spectral Techniques to Enhance the Intelligibility of Speech using a Harmonic Model”, Proc. of InterSpeech 2012 (2012)
2. S. King, J. Yamagishi, C. Valentini-Botinhao, “Speech intelligibility enhancement for HMM-based synthetic speech in noise”, Proc. SAPA-SCALE Conference (2012)
3. Cassia Valentini-Botinhao, Ranniery Maia, Junichi Yamagishi, Simon King and Heiga Zen, “Cepstral analysis based on the glimpse proportion measure for improving the intelligibility of hmm-based synthetic speech in noise”, Proc. of ICASSP 2012 (2012)
4. T. Raitio, A. Suni, H. Pulakka, M. Vainio and P. Alku, “Comparison of formant enhancement methods for HMM-based speech synthesis”, Proceedings the 7th ISCA Tutorial and Research Workshop on Speech Synthesis, 334-339 (2010)
5. G. K. Anumanchipalli, P. K. Muthukumar, U. Nallasamy, A. Parlikar, A. W. Black, B. Langner, “Improving speech synthesis for noisy environments”, The 7-th Speech Synthesis Workshop (2010)
6. A. Suni, T. Raitio, M. Vainio, and P. Alku, “The GlottHMM speech synthesis entry for Blizzard Challenge 2010,” in Proc. Blizzard Challenge Workshop 2010, Kyoto, Japan (2010)
7. Dong-Yan Huang, Susanto Rahardja and Ee Ping Ong, “Lombard Effect Mimicking”, The 7-th Speech Synthesis Workshop (2010)
8. Lane H, Tranel B. “The Lombard sign and the role of hearing in speech”. J Speech Hear Res 14 (4): 677–709 (1971)
9. HMM-based speech synthesis system (HTS), http://hts.sp.nitech.ac.jp/, accessed Nov. 30, 2012
10. K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, “Speech Parameter Generation Algorithms for HMM-based Speech Synthesis”, Proc. of ICASSP, 3, 1315-1318 (2000)
11. B.H. Juang and S. Katagiri, “Discriminative learning for minimum error classification”, IEEE Trans. Signal Processing, 40 (12), 3043-3054 (1992)
12. Y.J. Wu and R.H. Wang, “Minimum generation error training for HMM-based speech synthesis,” Proc. of ICASSP 2006, 1, 89-92 (2006)
13. S. King and V. Karaiskos, “The Blizzard Challenge 2009”, Blizzard Challenge Workshop 2009, http://www.festvox.org/blizzard/bc2009/summary_Blizzard2009.pdf, accessed Nov. 30, 2012
14. NOISEX-92 noise database, http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html, accessed Nov. 30, 2012
15. Y.F. Liao and M.L. Wu, “The NTUT Blizzard Challenge 2009 entry”, Blizzard Challenge 2009 Workshop, http://festvox.org/blizzard/bc2009/ntut_Blizzard2009.pdf, accessed Nov. 30, 2012