INTERSPEECH 2009 BRIGHTON

Unit Selection based Speech Synthesis for Poor Channel Condition

Ling Cen, Minghui Dong, Paul Chan, Haizhou Li

Institute for Infocomm Research (I^2R), A*STAR, 1 Fusionopolis Way, Singapore 138632

{lcen, mhdong, ychan, hli}@i2r.a-star.edu.sg

10.21437/Interspeech.2009-595

Abstract

Synthesized speech can be largely degraded in noise, resulting in compromised speech quality. In this paper, we propose a unit selection based speech synthesis system for better speech quality under poor channel conditions. First, the measurement of speech intelligibility is incorporated in the cost function as a searching criterion for unit selection. Next, the prosody of the selected units is modified according to the Lombard effect. Prosody modification includes increasing the amplitude of unvoiced phoneme and enlarging the speech duration. Finally, the FIR equalization via convex optimization is applied to reduce signal distortion due to the channel effect. Listening test in our experiments shows that the quality level of synthetic speech can be improved under poor channel conditions with the help of our proposed synthesis system.

Index Terms: speech synthesis, unit selection, unit costs, Speech Intelligibility Index, equalizer, convex optimization

1. Introduction

With the increasing applications for speech technology in the real world, speech synthesis has become a point of great interest. The synthesis of human speech enables machines to communicate with human beings, blind or sighted alike, in the way most natural to them. However, poor channel conditions can severely degrade the quality of synthesized speech, greatly affecting its intelligibility. Currently, most research in speech synthesis is focused around quality of the speech under normal conditions.

In [1], speech intelligibility is improved by increasing the amplitude of consonants relative to that of vowels. It is found that humans tend to increase the consonant-vowel energy ratio when asked to speak clearly [2]. According to the Lombard effect, speakers alter their vocal projection in noisy environments. Measurable differences in vowel duration and intensity can be observed when examining the acoustic differences between Lombard speech and normal speech [3]-[5]. Based on the observation of the human speaking style, the prosody modification method is employed to improve intelligibility by increasing phoneme amplitude, altering spectral shape, and lengthening phoneme duration in [5]. In [6], an approach to improve intelligibility of synthetic speech using *speech in noise* is proposed. They have shown that altering the presentation of synthetic speech in similar ways that are used by humans when they are speaking in poor channel conditions improves the understandability of synthetic speech [6]. In [7], the measurement of intelligibility is incorporated in the cost function for unit selection. The calculation of the intelligibility values is done offline with additive pink noise.

In this paper, we propose a unit selection speech synthesis system, which is developed for synthesizing speech in poor channel condition. The proposed system includes 3 processing modules, which is elaborated in Section 2. The listening test shown in Section 3 indicates that the quality level of synthetic speech can be improved in noisy environments using the proposed system. The concluding remarks are given in Section 4.

2. Proposed Unit Selection Speech Synthesis System

2.1. System architecture

The synthesized speech may suffer severe distortion under poor channel conditions. The major challenge here is that the intelligibility of speech can be severely affected due to channel noise. In this section, a unit selection based system is proposed for speech synthesis in poor channel conditions (for example, in telephonic channels).

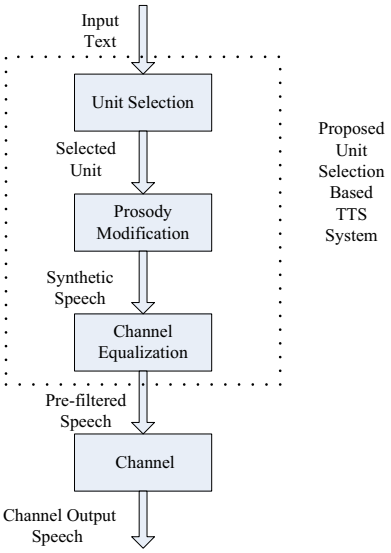

Figure 1: System flowchart for speech synthesis under poor channel condition.

There are 3 processing modules in the proposed Text-to-Speech (TTS) system, as shown in Fig. 1. These are the unit selection, prosody modification and channel equalization modules. In order to improve speech intelligibility, and hence, the quality of synthetic speech, the following methods are taken in our system.

(1) The speech intelligibility is measured and incorporated in the cost function as a searching criterion for unit selection. The intelligibility is evaluated using Speech Intelligibility Index (SII) [8]. Unlike the method in [7], we calculate the SII with the actual noise presented in the channel condition.

(2) The prosody of the generated speech is modified by increasing the intensity of unvoiced phonemes and lengthening the duration of speech before they are concatenated.

(3) In order to reduce signal distortion caused by the channel effect, FIR equalization is used to post-process the

synthetic speech, which is designed using convex optimization.

2.2. Intelligibility measurement in unit selection

2.2.1. Speech Intelligibility

One of the most important tasks in speech synthesis with poor channel conditions is to increase speech intelligibility in a noisy environment. It has been reported that the intelligibility of synthetic speech can be degraded in noise [6], [7]. To address the problem, an intelligibility cost is incorporated in the unit cost function as a searching criterion for unit selection. In this way, units with larger values for intelligibility tend to be selected.

In our work, the SII [8] is used to measure the intelligibility cost of each unit. The SII represents a physical measure that is highly correlated with the intelligibility of speech. It has been evaluated by speech perception tests given to a group of talkers and listeners. The SII is calculated from acoustical measurements of speech and noise, whose value varies from 0 (completely unintelligible) to 1 (perfect intelligibility).

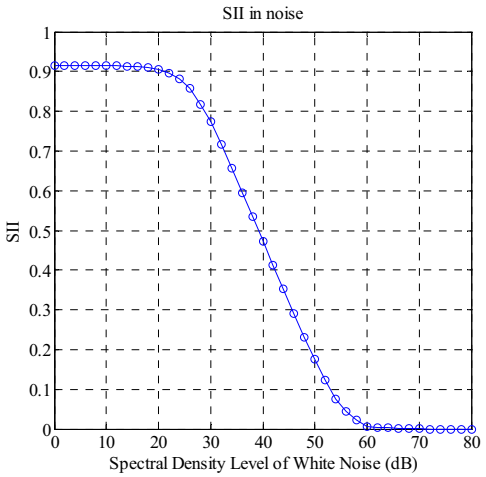

Figure 2: SII in noise with different density levels of white noise.

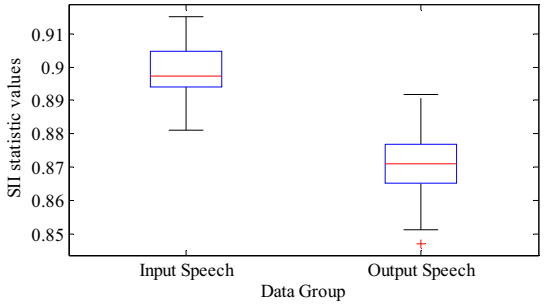

Figure 3: Channel effect on SII.

Figure 2 illustrates the SII measurement of speech intelligibility in various spectral density levels of white noise which is evaluated using the program given in [9]. The speech sample is processed by an eighteen-band third-octave filterbank. Multirat implementations of the filter is used in the range of 1250 Hz band to 100 Hz band and direct implementation is for the range of 5000 Hz to 1600 Hz. The power of the filterbank output in bands 1 through 18 is used in SII calculations. It can be seen from the figure that the SII values decrease with the increase of the noise levels.

The box-plot in Fig. 3 examines the effect of the telephony channel on speech intelligibility. This figure shows the SII values of 20 utterances before and after transmission through the telephone channel. This figure indicates that the channel noise has measurable influences on the speech intelligibility index.

2.2.2. Unit Selection

The unit selection process is based on the cost function that consists of three parts.

1) A target cost to measure the difference between the target unit and the candidate unit.

2) A join cost to measure the acoustic smoothness between the concatenated units.

3) An intelligibility cost to measure the predicted intelligibility of the candidate unit after passing through the channel.

The target cost further consists of two parts, i.e. the cost of acoustic parameters, and the cost of context linguistic features. The target cost c_t is defined as

$$c_t = w_{ta}c_{ta} + w_{tl}c_{tl} \quad (1)$$

where, c_{ta} and c_{tl} are costs of acoustic parameters and linguistic features, respectively, and w_{ta} and w_{tl} are the weighting parameters for c_{ta} and c_{tl}, respectively.

The join cost, c_j is defined as the squared value of the Euclidean distance between the vector of the end frame in the previous unit E_{i-1} and the vector of the start frame in the current unit S_i as

$$c_j = (E_{i-1} - S_i)(E_{i-1} - S_i)^T \quad (2)$$

In order to predict the intelligibility of each unit in the channel, we calculate the SII for the unit of channel output speech. The intelligibility cost c_s is defined as the following:

$$c_s = 1 - q \quad (3)$$

where q is the SII for the unit.

The total cost, c is defined as

$$c = w_t \sum_{i=0}^{n} c_t(i) + w_s \sum_{i=0}^{n} c_s(i) + w_j \sum_{i=1}^{n} c_j(i) \quad (4)$$

where n is number of units in the sequence, $c_t(i)$ is the target cost of unit i, $c_s(i)$ is the intelligibility cost of unit i, $c_j(i)$ is the join cost between units i-1 and i, and w_t, w_s and w_j are weights for target, speech intelligibility and join costs, respectively.

Besides the intelligibility cost in the cost function, several other measures are being taken to further improve the intelligibility by filtering out the units that are possibly poor intelligibility.

(1) Some of the units are not correctly uttered in the recording. Some are uttered very similar to other phonemes. The selection of these units will lead to unclear or wrong sounds. Therefore, we use speech recognition techniques to filter out these units. The generalized posterior probability (GPP)-based confidence measure method is used [10], [11]. The units that receive a poor log likelihood ratio (LLR) under a threshold are discarded from unit selection.

(2) In natural speech, some of the phonemes are weakened while speaking. These units are not as clear as other units. The weakened units normally have a shorter duration.

Therefore, filtering out the units that have very short duration helps to improve overall intelligibility.

2.3. Prosody modification in speech synthesis

First described by Etienne Lombard in 1911, the Lombard effect is a phenomenon in which speakers alter their vocal production in noisy environments. Measurable differences have been found in vowel duration and intensity in previous research examining the acoustic differences between Lombard speech and normal speech [3]-[4]. Prompted by the influence of the Lombard effect on the speakers, we modify the speech prosody, in order to improve the understandability of synthetic speech in poor channel conditions.

In our method, prosody modification involves increasing the intensity of unvoiced speech segments and lengthening the duration of speech without affecting its naturalness. The trade-off between speech naturalness and comprehensibility is considered when we choose the modification magnitude for increase of intensity and lengthening of duration.

In the unit selection TTS system, the selected units will be concatenated together to form the synthetic speech. As we know the phone type of each unit, it is easy for us to only modify the amplitude of the unvoiced part. When all the units are joined to form a speech utterance, we lengthen the utterance with a very high quality speech analysis/synthesis engine namely STRAIGHT [12]. The STRAIGHT system, based on the source filter model, allows flexible control of speech parameters. Its conceptual simplicity has made this system a tool for speech perception research as well as other speech research applications.

2.4. Channel effect reduction using FIR equalization

When the synthetic speech is transmitted via a channel (e.g. telephony), the quality of the speech received is degraded due to the channel effect. Channel equalization technology is employed to reduce such an effect. This is elaborated below.

Assume that the speech signal $s(t)$ is transmitted from node A to node B through a channel. Let the received speech signal at node B be $y(t)$ and the time-domain impulse response of the channel be $h(t)$. There is

$$y(t) = (s * h)(t). \tag{4}$$

In order to remove the channel effect in the received speech at node B, we can design an equalizer with impulse response of $g(t)$, which satisfies

$$(g * h)(t) = \delta(t). \tag{5}$$

where $\delta(t)$ is the impulse signal. The synthesized speech is first filtered by $g(t)$ before transmission. The signal received at Node B, $\hat{y}(t)$ can be expressed as

$$\hat{y}(t) = (s * g * h)(t). \tag{6}$$

Theoretically, if (5) can be accurately realized, the received speech $\hat{y}(t)$ will have no distortion due the channel effects and we can have

$$\hat{y}(t) = s(t). \tag{7}$$

The channel effect may be suppressed this way in practice, to some extent, depending on the performance of the equalizer.

Equalizer design is a typical problem in signal processing. In our method, convex optimization method is employed. A convex optimization problem is the minimization of a convex function over a convex set. A set is convex if the line joining any pair of its points lies within the set. A function ϕ, is convex on a convex set if it satisfies the condition [13] as

$$\phi(\lambda x + (1-\lambda)y) \le \lambda\phi(x) + (1-\lambda)\phi(y), 0 \le \lambda \le 1. \tag{8}$$

In convex optimization, any local minimum of a convex function is a global minimum. Convex optimization is an excellent tradeoff in efficiency between the analytical methods and the numerical techniques [13].

The equalizer is realized as an FIR filter, which is designed based on the Chebychev criteria. The problem can be specified in terms of frequency response functions as

$$\min_{g} \max_{w\in[0,\pi]} |H(w)G(w) - H_d(w)|, \tag{9}$$

where $H(w)$ and $G(w)$ are the frequency responses of the channel and the equalizer, respectively, and $H_d(w)$ is the desired frequency response. This can be formulated as a Second Order Cone Programming (SOCP) and solved efficiently using convex programming.

3. Experiments

Our experiments are carried out based on the telephone channel. For telephone channel, the synthesized speech is firstly down-sampled to 8k Hz before transmission. In our work, the telephone channel is simulated by the telephone channel simulation tool provided by the organizers of Blizzard Challenge to its participants. Whenever a speech utterance is generated, we perform listening tests after the processing of the simulation tool.

The speech corpus that we use in this research is the British English corpus provided by the University of Edinburgh for Blizzard Challenge 2009. The released part of the corpus consists of 15 hours speech in 9,509 utterances, which cover children stories, isolated words, sentences carrying emphasis, news articles, etc [14], [15]. The corpus is designed to cover as much variations of diphones as possible.

3.1. Prosody modification

In prosody modification, the duration of the speech is enlarged to improve the speech quality. We need to decide how much we should lengthen the duration of the speech. Considering improving the speech intelligibility, the unit selection process favours longer units. Hence, the duration of the synthesized speech tends to be longer. We have generated the speech with different lengthening rates, i.e. 1.1, 1.2, 1.3, and 1.4. After carefully examining the generated speech and comparing it with the speech without lengthening, it is found that a moderate lengthening rate of 1.2 is a good choice without compensating the naturalness.

We have observed that the unvoiced part of speech can be degraded largely when the speech samples are transmitted via the telephone channel. Therefore, it is necessary to increase its amplitude, in order to preserve the intelligibility of unvoiced parts. As our TTS system is a unit selection based system, we know the exact unit composition of each segment when generating the speech samples. This makes it easy to identify the unvoiced unit without using an unvoiced detection program.

Forty sentences are generated for testing. A sequence of units is selected for each sentence by the unit selection process. Before concatenation, the amplitude of unvoiced part is increased by a rate ranging between 1.1 and 2.0. The speech utterances are lengthened to 1.2 times as long. After the generated speech samples are processed using the telephone channel tool, the average SII values for the utterances are calculated. The average SII values for the set of utterances in different amplifying rate are as shown in Figure 2. In this figure, the *x* axis represents the amplification factor for the unvoiced parts of the speech signal. The *y* axis shows the SII values. From the figure, it can be seen that the speech intelligibility increases with an increase in the amplification factor (even though this is not significant). This shows that increasing the amplitude of the unvoiced part helps to improve speech intelligibility.

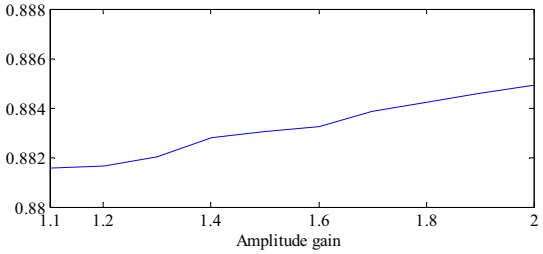

Figure 4: Speech intelligibility with the increase in amplification factor.

3.2. Subjective listening test

To test how the proposed method helps to improve speech quality, a listening test is conducted on 40 testing sentences (20 news style and 20 conversation style). We compare the quality of synthesized speech based on two different methods:

- Method A: We disable all other measures used to improve intelligibility for poor channel conditions. That is, we give the weight of intelligibility a value of zero, not use the prosody modification approach and not apply channel equalization.
- Method B: We apply all the measures mentioned above and set the amplification factor for the amplitude of the unvoiced speech to 2 times that of the original signal. Also we set the duration to be lengthened to 1.2 times of the original signal.

Table 1. *Result of listening test*

Method	A	B
Percentage of listeners' preference	32%	68%

We synthesize 40 utterances with the 2 different methods. The two sets of utterances are then passed through the telephone channel simulator. A listening test involving 10 listeners is conducted. In the test, we ask each listener to tell which utterance in the pair is better. The result is as shown in Table 1. From the table, we see that 68% of the 400 votes prefer the utterances that are synthesized based on method B. Therefore, it is indicated that the use of the proposed TTS system helps to improve the quality of speech in poor channel conditions.

4. Conclusions

The paper proposes a unit selection based speech synthesis system to improve the quality of speech synthesized for poor channel conditions. The improvement is achieved through the implementation of three methods. The integration of intelligibility cost into the cost function to select the units with higher intelligibility, application of the prosody modification process to enhance the intelligibility of the unvoiced segments in speech and the application of FIR equalization via convex optimization to reduce the channel effect. Listening tests show that our system can generate better quality speech than that without adopting these methods.

5. References

[1] E. A. Kretsinger, and N. B. Young, "The Use of Fast Limiting to Improve the Intelligibility of Speech in Noise," *Speech Monogr.*, no. 27, pp. 63-69, 1960.

[2] M.A. Pichney, N.I. Durlach, and L. D.Braida, "Speaking Clearly for the Hard of Hearing II: Acoustic Characteristics of Clear and Conversational Speech," *J. Speech and Hearing Research*, vol. 29, pp. 434-446, 1986.

[3] Junqua, Jean-Claude, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," *Speech Communication*, vol. 20, pp. 13-22, 1996.

[4] Summers, W. Van et al, "Effects of noise on speech production: Acoustic and perceptual analyses," *J. Acoust. Soc. Am.*, vol. 84, no. 3, pp. 917-928, 1988.

[5] D. Pan, B. Heng, S. Cheung, and E. Chang, "Improving Speech Synthesis for High Intelligibility under Adverse Conditions," in *6th International Conference on Spoken Language Processing*, Beijing, Oct. 2000.

[6] B. Langner, and A. W. Black, "Improving the Understandability of Speech Synthesis by Modeling Speech in Noise," in *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 265-268, Philadelphia, 2005.

[7] MiloŠ Cerňak, "Unit Selection Speech Synthesis in Noise," in *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 761-764, France, 2006.

[8] ANSI-S3.5, "American National Standard, Methods for Calculation of the Speech Intelligibility Index," 1997.

[9] H. Muesch, "SII: Speech Intelligibility Inde, http://www.sii.to/, a Matlab program," 2005.

[10] L. Wang, Y. Zhao, M. Chu, F. K. Soong, and Z. Cao, "Phonetic transcription verification with generalized posterior probability," in *Eurospeech*, pp. 1949-1952, 2005,

[11] Z. Ling, H. Lu, G. Hu, L. Dai, and R. Wang, "The USTC system for Blizzard Challenge 2008," in *Blizzard workshop*, 2008.

[12] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction,'' *Speech Commun*., vol. 27, pp. 187–207, 1999.

[13] Hervé Lebret, and Stephen Boyd, 'Antenna Array Pattern Synthesis via Convex Optimization', *IEEE Transactions on Signal Processing*, vol. 45, no. 3, pp. 526-532, March 1997.

[14] V. Strom, R. Clark, and S. King, "Expressive Prosody for Unit-Selection Speech Synthesis," in *Proc. Interspeech*, Pittsburgh, 2006.

[15] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, and D. Jurafsky, "Modelling Prominence and Emphasis Improves Unit-Selection Synthesis," in *Proc. Interspeech*, Antwerp, 2007.