ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Objective Intelligibility Assessment of Text-to-Speech System using Template Constrained Generalized Posterior Probability

Linfang Wang[1], *Lijuan Wang*[2], *Yan Teng*[1], *Zhe Geng*[1], *Frank K. Soong*[2]

[1]Microsoft Bing, [2]Microsoft Research Asia, Beijing, China

{linfangw, lijuanw, yanteng, zhegeng, frankkps}@microsoft.com

10.21437/Interspeech.2012-195

Abstract

Speech intelligibility is one of the most important measures in evaluating text-to-speech (TTS) synthesizer. In this paper, we propose an automatic objective intelligibility measure for evaluating synthesized speech using template constrained generalized posterior probability (TCGPP). TCGPP is a posterior probability based confidence measure, which has the advantage to identify small granularity errors in synthesized speech. Moreover, the TCGPP scores over a test set can be summarized into an overall objective intelligibility score to compare two synthesizers, or rank multiple TTS systems. We conducted the experiments using the synthesized test sentences from all the participants of EH1 English task in Blizzard Challenge 2010. The results show the proposed measure has high correlation ($corr = 0.9$) with subjective scores and ranking.

Index Terms: speech synthesis, objective intelligibility, Template constrained generalized posterior probability

1. Introduction

Speech intelligibility is an overall perception of whether the speech being heard exactly matches what the speaker is trying to convey. Speech intelligibility is one of the most important speech quality measurements for speech coding, speech synthesis and speech enhancement applications. Thus, speech intelligibility evaluation is of primary interest for these areas.

In most cases, speech intelligibility is subjectively measured by human. Typically, language experts or native speakers are hired to take yes/no understandable checking tests or dictation tests. Average understanding rate or word error rate is calculated as the subjective speech intelligibility score for a particular speech coding or synthesis application. However, human listening tests are often costly, labor intensive, prone to inconsistency among subjects, and hard to scale up for a large comprehensive test. The need for automatic objective intelligibility evaluation is continuously growing. In particular, for developing applicable TTS systems, automatic objective intelligibility assessment is desired for voice quality comparison/benchmark and contribution measurement of various component approaches.

The probably earliest attempt for objective intelligibility evaluation dates back to 1947 when Bell Labs developed the articulation index (AI) [1]. Several variations based on the AI have been developed, including the speech transmission index (STI) which is included in IEC standard 60268-16 [2]. Both AI and STI correlate well with subjective intelligibility scores but their applicability is rather limited to linear systems, and not quite suited to modern applications. In 1999, Chernick et al [3] proposed to use speech recognizers for automated intelligibility scoring, and reported good correlation with subjective scoring for CELP-encoded speech samples. In 2002, Jiang et al [4] investigated automatic speech recognition (ASR) in the context of the G.729 codec and also reported promising results.

However, all above findings work are in the context of speech coding, not necessarily applicable for TTS systems. In 2008, Vích et al [6] proposed using word recognition rate of an ASR system to do intelligibility assessment for Czech TTS systems. However, the correlation between the speech recognition results and subjective test results was not shown in the study. In 2009, Cerňak et al [7] did similar study for Slovak language, and presented promising results in terms of correlation between the subjective intelligibility scores and the ASR scores calculated for their own male unit-selection voice. These two studies show that speech recognition based technologies have the potential for objective intelligibility evaluation of TTS systems.

In our previous work [8][9][10], the TCGPP(Template Constrained Generalized Posterior Probability) has been proposed and demonstrated its effectiveness for speech transcription verification. It is a natural thought that this algorithm may also be helpful in identifying unintelligible speech synthesis errors produced by TTS systems.

In this paper, we propose an objective intelligibility evaluation method for synthesized speech based upon the TCGPP algorithm. Experiments on the test data of Blizzard Challenge 2010 EH1 task show that, the proposed objective intelligibility measure correlates well with subjective intelligibility scores for all the entries with difference synthesis approaches including the original recordings.

The rest of the paper is organized as follows. Section 2 briefly introduces the TCGPP algorithm. Section 3 presents the proposed objective intelligibility evaluation method for synthesized speech. Section 4 gives the experimental results and analysis. Section 5 draws the conclusions.

2. Template Constrained Generalized Posterior Probability (TCGPP)

TCGPP is an extension of the generalized posterior probability (GPP) [9,10]. Since the templates are flexibly constructed, TCGPP can either be reduced to the traditional GPP, which considers only the focus unit, or be built upon a template of complex topology, where specific context for the focus word is defined. Moreover, the templates allows a "sifting" of hypotheses; only those hypotheses that match both the focus unit and the specified contexts are included in the search space, which leads to higher calculated probability for the focus unit and greater confidence.

2.1. Template and its variation

We denote a Template by a triple [$\mathcal{T}$; *r; s, t*]. Template $\mathcal{T}$ is a pattern composed of hypothesized units and metacharacters that can support regular expression syntax; *r* stands for the partial match ratio and ranges between 0 and 100%. This means the relevant path needn't 100% match the template. [s, t] defines the time frame constraint on the template.

As shown in Fig. 1, basic template T_1 depicts the simplest type of template, ABCDE, where C is the focus unit, and AB and DE are the left and right context respectively. Template T_2, A*CDE, includes a wild-card * that indicates an arbitrary character in that particular position: A*CDE matches AACDE, AFCDE, or ACDE. Template T_3, ABCϕE, includes a blank, ϕ, to indicate a null in this position. Template T_4, ABC?E, includes a question mark, ?, to indicate that the word which appears in this position has not been identified yet.

These basic templates can be combined to construct a compound template, such as template T_5 depicted in Fig. 1. With reference to compound template T_5, a matching string hypothesis may include either A or K in the 1st position, include B or any element at the 2nd position, includes C at the center position, and so on. Depending upon the specified minimal matching constraint and whether some or all of these elements can be partially matched, the search space generated from compound template T_5 may be substantially larger than that generated from a basic template.

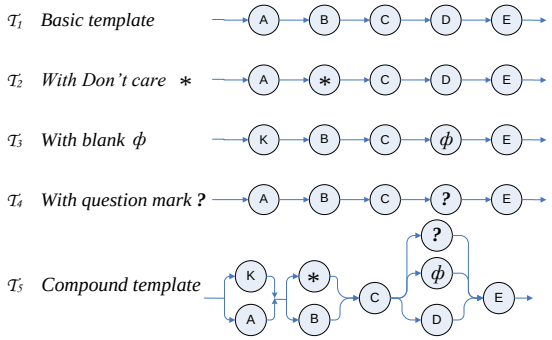

Fig. 1: *Illustration of templates*

2.2. TCGPP calculation

Once a template is constructed, an appropriate hypothesis set H([$\mathcal{T}$; *r; s, t*]) is determined by matching all the string hypothesis against the template. The hypothesis set under stringent template constraints can be much smaller than that under the traditional GPP approaches. The Template Constrained Posterior (TCGPP) of [$\mathcal{T}$; *r; s, t*] is calculated as the generalized posterior probability summed on all the string hypotheses in H([$\mathcal{T}$; *r; s, t*]), as Eq. 1 shows.

$$P([\mathcal{T};r;s,t]\,|\,x_1^T) = \sum_{\substack{N,h=[w,s,t]_1^N \\ h\in H([\mathcal{T};r;s,t])}} \frac{\prod_{n=1}^{N} p^{\alpha}(x_{s_n}^{t_n}\,|\,w_n)\cdot p^{\beta}(w_n\,|\,w_1^N)}{p(x_1^T)} \quad (1)$$

where x_1^T is the whole sequence of acoustic observations, α and β are the exponential weights for the acoustic and language model likelihoods, respectively. In calculating TCGPP, the reduced search space, the time relaxation registration, and the weighted acoustic and language model likelihood are handled similarly as in GPP [2]. The difference between the TCGPP and GPP is the determination of the string hypotheses set, which corresponds to the term under the sigma summation notation.

The TCGPP approach examines both the focused unit and the context to the left and right of the focused unit. In this way, the TCGPP approach provides additional robustness against incorrect time boundaries, which may be caused by insertion, deletion, or substitution errors [6]. Also, the proposed template constrained approach uses templates to limit the hypothesis set during the posterior probability calculation for a selected focus unit. These templates may be tailored to different granularity. This makes it possible to measure confidence at different precision levels.

3. Objective Intelligibility Evaluation of Synthesized Speech using TCGPP

Here we are using TCGPP to evaluate the synthesized speech against the text we feed into the synthesizer. The challenge is that given the state-of-the-art speech synthesis technology, the synthesized speech already achieves a relatively high intelligibility and naturalness. That means, the errors in synthesized speech is rare, and at very small granularity, for example, a weak reproduce of one or half phoneme. The flexible property allows TCGPP to examine the synthesized speech at fine granularity level, given the input text transcription as reference. To achieve that, multiple templates are constructed with loose to restrict configurations and multiple confidence scores can be obtained for each individual phone. Moreover, an overall speech intelligibility score can be further calculated based on local intelligibility confidence scores to compare and rank multiple TTS systems.

3.1. TCGPP of a focused phone

Phone level TCGPP is used as the confidence measure to identify potential phone errors in synthesized speech sentences. A template [$\mathcal{T}$; *r; s,t*] for a focused phone is constructed as shown in Fig. 2. p_k is the focused phone, $p_{k-L}\ldots p_k\ldots p_{k+L}$ is the phone string covering the 2*L* context phones before and after p_k. $\tilde{p}_k$ represents the confusable phone of p_i (k-L≤i≤k+L). The confusability between two phones is assessed by the Kullback-Leibler Divergence (KLD), which is a measure of the dissimilarity between two probabilistic models. *r* is the partial match ratio among the *2L* context phones. [s, t] defines the time frame constraint of the template, i.e., *s* is the start time of p_{k-L} and *t* is the end time of p_{k+L}. The correct hypotheses set H for [$\mathcal{T}$; *r; s, t*], as defined in Eq. 1, is obtained by finding every string hypothesis that contains a subpath that r% partially matches the template and also overlaps the specified time interval [s, t].

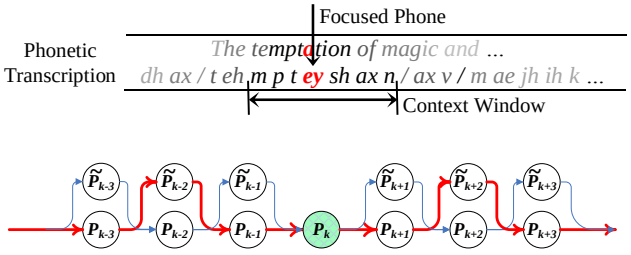

Fig. 2: *Illustration of a focused phone p_k, and its template*

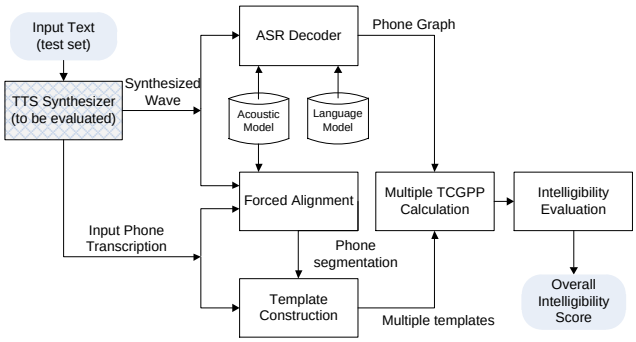

Fig. 3: *A flowchart of objective intelligibility assessment procedure for a TTS system.*

3.2. Objective intelligibility assessment of TTS system

Fig.3 shows the flowchart of the objective intelligibility evaluation procedure for a TTS system, which can be accomplished in following steps.

Step 1. Test speech synthesis. Firstly, for a TTS synthesizer to be evaluated, text script of a test set is feed as input. The correct phone transcription of the test set is also needed to provided, which can be generated by a standard frontend and then refined by human experts to avoid frontend errors. The output is the synthesized waveforms for this test set.

Step 2. Phone graph decoding. With acoustic model and language model, ASR phone decoder generates phone graph for a spoken input. The acoustic model can be trained speaker independently or dependently. The language model used in the decoder is phone tri-gram model.

Step 3. Forced alignment. In order to get the starting/ending time boundaries for each phoneme, forced alignment is carried out between the initial phone transcription and the acoustic signals. The acoustic model is the same one used in phone graph decoding.

Step 4. Confusable phone pairs generation. The confusability of each phone pair is evaluated by KLD calculated upon the acoustic model.

Step 5. Template construction and TCGPP calculation. Each phone in the initial phone transcription is regarded as a focused phone, for example the focused phone "ey" in Fig. 2. Rather than construct one template with optimized parameters [9], we construct multiple templates according to the focused phone and its left and right context phones. By setting the context window length, the threshold for selecting the number of confusable phones, and the partial matching ratio, multiple templates are generated. Some more rigid templates are constructed according to the specific context, while others are more flexibly constructed with more confusable phones or lower partial match ratio. The motivation is to use multiple templates of different hypothesis sifting capabilities -- from well-defined, full context to loosely defined context like wild card to measure corresponding confidence at different expected accuracy. The TCGPP values for all the templates of each phone are calculated.

Step 6. Intelligibility evaluation. Once the TCGPP calculation is complete, we can start to evaluate objective intelligibility. For each focused phone in the input phonetic transcription, we calculate multiple TCGPP values to evaluate the focused phone at different precision. The syllable and/or word level confidence score is calculated by averaging or taking the minimum of the confidence scores of the phones it contains. The scores for phone, syllable, and word are all posterior probabilities, ranging from 0 to 1. Next, we can summarize these posterior scores into a single metric to represent the intelligibility of the TTS system.

3.3. Overall speech intelligibility score calculation

For each focused phone in the transcription, we calculate multiple TCGPP values. Basically, the overall speech intelligibility score can be calculated based on the various confidence scores calculated above. The overall objective intelligibility score (OIS) can be calculated as follows:

$$\mathrm{OIS} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \omega_{\mathrm{k}} \left(\frac{1}{\mathrm{M}_k} \sum_{\mathrm{m}=1}^{\mathrm{M}_k} \mathrm{P}_{\mathrm{k,m}} \right) \quad (2)$$

where $\mathrm{k} = 1, \ldots, \mathrm{K}$ represents the different granularity levels (i.e., phone, syllable, word, utterance, etc.) that are taken into account; ω_{k} is the weight for the k^{th} level; $\mathrm{m} = 1, \ldots, \mathrm{M}_k$ represents the number of focused units at the k^{th} granularity level; $\mathrm{P}_{\mathrm{k,m}}$ is the average of the multiple TCGPP confidence scores for the focused unit. The raw objective speech intelligibility can be linear transformed to any specific range, e.g., [0,100].

4. Experimental Results

4.1. Experimental setup

We use the proposed framework to evaluate different TTS voices. As a reference, word error rate (WER) by human dictation on a test set is taken as the ground truth of the voice intelligibility. Meanwhile, the accuracy of ASR, word recognition rate (WRR), is employed as the baseline of the objective intelligibility assessed by computer. The proposed objective intelligibility measure can be compared with both human (upper bound) and ASR (lower bound) results.

The first experiment is done on comparing different versions of our Microsoft internal TTS systems. The second experiment is more comprehensive, using the test data of all the entries in Blizzard Challenge 2010 EH1 task.

4.2. Experiment on comparing internal voice fonts

Three voice fonts are built by three different TTS systems using 6000 utterances read by a female US native speaker. These three TTS voices are built with different backend synthesis technologies: two (voice1 and voice2) are unit selection based; and the third one (voice3) is HMM based synthesis. As the three systems use the same frontend, this experiment is mainly to compare different backend technologies. A test set of 330 normal text sentences is generated by the three TTS systems respectively. Their speech intelligibility is evaluated by human language experts by examining the mispronounced word after listening to the synthesized speech repeatedly. Also, automatic speech recognition is conducted on the same test set to get the WRRs as the baseline for objective intelligibility assessment.

Please note that the word error rate (WER) by human listening test is converted to subjective understanding rate (SUR) to make them congruent with intelligibility score we calculated:

$$\mathrm{SUR} = 100 - \mathrm{WER} \quad (3)$$

Table 1 shows the SUR (by human), the WRR (by ASR), and the proposed objective overall intelligibility scores (linearly normalized to [0, 100]). In this task, both the ASR and the

proposed measure get the same rank ordering as that of human perception.

Table 1: *Subjective and objective intelligibility score of the three TTS voice fonts.*

TTS voice	SUR (by human)	WRR (by ASR)	Objective Intelligibility Score
1	98.4	58.2	59.3
2	99.1	68.5	61.8
3	99.5	73.0	67.6

4.3. Experiment on Blizzard Challenge 2010 EH1 task

Another experiment is done on the test data of the 17 entries in the Blizzard Challenge 2010 English EH1 task [11]. All these 17 systems are trained on 4014 sentences, by a British English male voice. Blizzard Challenge had speech experts, native speakers, and other volunteers to take dictation tests on 100 sentences in the SUS (Semantically Unpredictable Sentences) session, and calculate the average WER for each system. The system with the lowest WER got the first place in intelligibility evaluation.

Fig.4 shows the three resultant curves: the SUR by human (reference), the WRR by ASR (baseline), and the objective intelligibility scores calculated by the proposed framework. The natural speech is represented by letter *A*. The proposed objective intelligibility is generally in line with the subjective scores with just a few outliers. In terms of the correlation coefficient with human intelligibility scores, the proposed measure achieves 0.9, which is much higher than 0.71 got by the accuracy of ASR.

5. Conclusions

We propose an objective intelligibility measure for assessing TTS systems using template constrained generalized posterior probability (TCGPP). TCGPP has the advantage to identify the errors in synthesized speech at small granularity level. Moreover, the TCGPP scores over a same testing set can be summarized into an overall objective intelligibility metric to compare two synthesizers, or rank multiple TTS systems. The experiments on Blizzard Challenge 2010 EH1 English task show that the proposed measure can estimate the objective intelligibility with high correlation to the subjective scores and ranking.

6. References

[1] N. R. French, J. C. Steinberg, "Factors Governing The Intelligibility of Speech Sounds," JASA, vol. 19, no. 1, pp. 90-119, 1947.

[2] H. J. M. Steeneken, T. Houtgast, "The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility," Acustica, vol. 28, pp. 66-73, 1973.

[3] C. M. Chernick, S. Leigh, K. L. Mills, and R. Toense, "Testing the Ability of Speech Recognizers to Measure the Effectiveness of Encoding Algorithms for Digital Speech Transmission," IEEE Int. Military Comm. Conf. (MILCOM), 1999.

[4] W. Jiang, H. Schulzrinne, "Speech Recognition Performance as an Effective Perceived Quality Predictor," IEEE Int. Workshop on Quality of Service, pp. 269-275, 2002.

[5] W. M. Liu, K. A. Jellyman, J. S. D. Mason, N. W. D. Evans, "Assessment of Objective Quality Measures for Speech Intelligibility Estimation," in Proc. ICASSP2006, Toulouse, France, 2006.

[6] R. Vích, J. Nouza and M. Vondra, "Automatic Speech Recognition Used for Intelligibility Assessment of Text-to-Speech Systems", Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction, pp. 136 – 148, Springer 2008.

[7] M. Cerňak, M. Rusko and M. Trnka, "Diagnostic Evaluation of Synthetic Speech Using Speech Recognition," in Proc. ICSV16, Kraków, Poland, 2009.

[8] L.J. Wang, Y. Zhao, M. Chu, F.K. Soong, and Z.G. Cao, "Phonetic Transcription Verification With Generalized Posterior Probability," in Proc. INTERSPEECH-2005, Lisbon, 2005.

[9] H. Zhang, L.J. Wang, F.K. Soong, "Context Constrained-Generalized Posterior Probability for Verifying Phone Transcriptions," in Proc. INTERSPEECH-2007, Antwerp, 2007.

[10] L.J. Wang, T. Hu and F. Soong, "Template Constrained Posterior For Verifying Phone Transcriptions," in Proc. ICASSP2008, Las Vegas, U.S.A., 2008.

[11] S. King, V. Karaiskos, "The Blizzard Challenge 2010," in Proc. The Blizzard Challenge 2010 workshop, Japan, Sept. 2010.

Fig. 4: *Subjective and objective intelligibility score of all the entries in Blizzard Challenge 2010 EH1 task. (Natural speech is letter A.)*