UNIT SELECTION SPEECH SYNTHESIS IN NOISE

Miloš Cerňak

Institute of Informatics, Slovak Academy of Sciences
Dúbravská 9, 84507 Bratislava Slovakia
Milos.Cernak@savba.sk

ABSTRACT

The paper presents an approach to unit selection speech synthesis in noise. The approach is based on a modification of the speech synthesis method originally published in [1], where the distance of a candidate unit from its cluster center is used as the unit selection cost. We found out that using an additional measure evaluating intelligibility for the unit cost may improve the overall understandability of speech in noise. The measure we have chosen for prediction of speech intelligibility in noise is Speech Intelligibility Index (SII). While the calculation of the SII value for each unit in the speech corpus was made off-line, a pink noise was used as a representative noise for the calculation. Listening tests imply that such a simple modification of the unit cost in unit selection synthesis can improve understandability of speech delivered under poor channel conditions.

1. INTRODUCTION

The idea of speech synthesis in noise is to produce synthesized speech understandable in poor channel conditions. Although there exists an industry claim (included in advertising texts and web sites of TTS products) that the current TTS products are adequate for noisy environments typically prevalent in automobiles, airports, offices and classrooms, recent study on intelligibility of AT&T NextGen, Festival, and IBM ViaVoice do not support it [2]. This study shows that the mentioned modern TTS products with total error rates ranging from 15.27% to 17.06% did slightly worse in noise than the best TTS products of the past (DECtalk 1.8). The study further revealed a surprising finding that noise affected all TTS products equally. No TTS technique was better suited to resist noise interference than others.

The background noise is known to have strong influence on the speaker and such a speech is modified by so called **Lombard effect**. Duration and intensity of the phonemes is changed, changes in the spectral domain can be observed. This fact was used in [3], where the authors proposed a speech synthesis modification based on the Lombard effect. The goal was to enhance intelligibility of synthetic speech presented via telephone in a noisy environment. Using the similar assumption about the Lombard effect, a special speech database for synthetic speech in noise was created [4] and used [5] recently. However, the authors reported degradation of intelligibility of speech for unit selection speech synthesis in noise.

In this paper we try to overcome these problems for unit selection speech synthesis, considering some segmental properties of units. We incorporate a prediction of intelligibility of the units in the unit cost. The intention is to choose more intelligible units from the speech database for the synthesis. Section 2 describes the method of unit selection speech synthesis in noise. Section 3 describes some experiments and the evaluation of the synthesis using listening tests. Finally, Section 4 briefly concludes this paper, and gives some future directions of the next research.

2. UNIT SELECTION SYNTHESIS

Having good experiences with doing speech synthesis in TTS-BOX [6], we implemented new Slovak unit selection TTS system in the similar framework, based on automatic clustering of similar units [1]. The units were clustered within a unit type (particular phone) according to their prosodic and phonetic context. A decision tree was constructed for each phone in the database using contextual features listed in Table 1. `Wagoon` tool of the Edinburgh Speech Library was used as the implementation of the Classification and Regression Tree (CART) method.

The appropriate decision tree was used to find the best cluster of candidate units for each target at synthesis time. The clusters form a lattice, which is weighted by the unit and concatenation costs (see Section 2.1). The distance of a candidate unit from its center forms the unit cost, and the difference of F0 plus Euclidean distance between pitch synchronous mel frequency coefficients of the two neighboring frames of concatenated units forms the concatenative cost. A search is then done to find an optimal path through the lattice (the best sequence of the units).

The author is now with the Institut Eurécom, Department of Communications Multimedia, 2229 route de Crêtes - B.P. 193, 06904 Sophia-Antipolis, France. E-mail: *Milos.Cernak@eurecom.fr.*

Features	Values
vowel or consonant	vowel consonant
vowel length	short long diphthong schwa
vowel height	high mid low
vowel frontness	front mid back
lip rounding	yes no
consonant type	stop fricative affricative nasal liquid
place of articulation	labial alveolar palatal labio-dental dental velar
consonant voicing	yes no

Table 1. Segmental features used in decision trees.

2.1. Method

We assume that the speech corpus is rich enough to have more realizations of the unit within one cluster with different intelligibility. The distance of the unit θ to its cluster center was taken for the unit cost. However, we found out that using one additional measure for the unit cost, related to the intelligibility of the unit, can improve understandability of synthetic speech in noise.

The intelligibility testing of synthetic speech in noise has shown listeners' sensitivity to small but structural sensitive changes in phonetic realization of speech segments [7]. We used the Speech Intelligibility Index (SII) to calculate intelligibility of speech segments in the presence of additive noise. The SII represents a physical measure that is highly correlated with the intelligibility of speech as evaluated by speech perception tests given a group of talkers and listeners. The SII is calculated from acoustical measurements of speech and noise. The value of the SII varies from 0 (completely unintelligible) to 1 (perfect intelligibility) [8]. Next section 2.1.1 describes calculation of SII values in detail.

2.1.1. Calculation of the SII values

As stated in the introduction, the idea of speech synthesis in noise is to elicit synthesized speech when playing it in poor channel conditions. As it was decided to make the calculation of the SII values off-line, a pink noise was used as a representative noise for the calculation. This decision resulted from our assumption that the stationary pink noise can serve as a first approximation of the non-stationary noises used in this research. It was taken from SpEAR database [9]. According to the authors the noise was originally acquired by sampling high-quality analog noise generator (Wandel & Goltermann). The noise exhibits equal power per 1/3 octave of the spectrum.

For each speech segment (unit) in the speech database, and a selection of the pink noise with the same duration, the speech intelligibility index was calculated. Speech/noise units were analyzed by one-third-octave filterbank with band center frequencies given in Hertz [160, 200, 250, 315, 400, 500, 630, 800, 1000, 1250, 1600, 2000, 2500, 3150, 4000, 5000, 6300, 8000]. For the range from 160 Hz to 1250 Hz, multirate filter implementation was used.

The power of all the speech and noise files was normalized with respect to the average power of the plain speech recordings. The audio industry doesn't have any standard defining comfortable listening level, but the movie industry has been working with the 83dB SPL standard for years (measured using a C-weighted, slow averaging SPL meter). It represents a comfortable average listening level, determined by audio-engineering professionals and resulting from years of listening experience. In order to normalize the power of the speech and noise files to 83 dB SPL, we used a single channel calibration pink noise signal given by SMPTE standard. For each of 18 analysed bands, decibel values were calibrated (or normalized) using the following formula:

$$P_c = 83 - (P_{ref} - P)\,, \tag{1}$$

where P_{ref} is calculated power for calibration pink noise signal given in dB FS (digital full scale), and P is a power of the output of a filter from one-third-octave filterbank given also in dB FS. The difference of the calibration power and the calibrated power tells us how much we need to scale the signal in order to make it average 83dB. To yield actual power for SII calculation, a value of 83 dB is then added.

Finally, normalized decibel values P_c were used as input to the implementation of SII calculation [10]. The intelligibility values for all units were then written into the catalogue of units, and used later during unit selection.

2.1.2. Optimal sequence selection

The search through the lattice has been already formulized in [11]. Let θ be a unit of the speech database. Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$ be a concatenation of N units. An overall cost between the units Θ and the targets T can be expressed as:

$$d\,(\Theta, T) = \sum_{j=1}^{N} d_u\,(\theta_j, t_j) + \sum_{j=1}^{N-1} d_c\,(\theta_j, \theta_{j+1}) \tag{2}$$

where $d_u\,(\theta_j, t_j)$ is a unit cost of the unit θ_j and t_j, calculated as a distance of a candidate unit θ_j to its cluster center, and $d_c\,(\theta_j, \theta_{j+1})$ is a concatenation cost of concatenated units θ_j and θ_{j+1}. The optimal sequence of units $\widehat{\Theta}$ can be found as the one that minimizes the overall cost:

$$\widehat{\Theta} = \arg\min_{\Theta} d\,(\Theta, T) \tag{3}$$

over all sequences in the lattice. We have used a Viterbi method to make the search efficient. The definition of $d_u\,(\theta_j, t_j)$ can be further decomposed into partial costs:

$$d_u\,(\theta_j, t_j) = \sum_{i=1}^{I} w_i d_u^i\,(\theta_j, t_j) \tag{4}$$

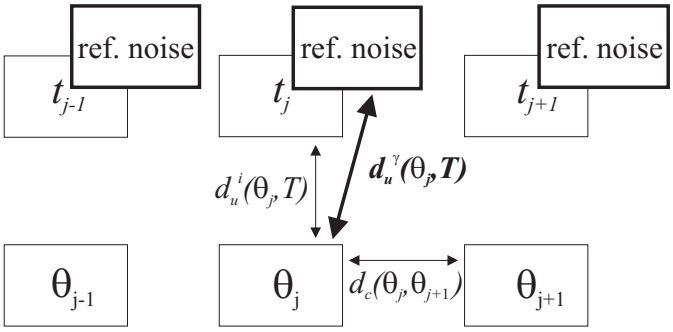

Fig. 1. The selection model of speech synthesis, where θ_j are candidate units, t_j are targets, $d_u = ()$ is the unit cost and $d_c = ()$ is the concatenation cost.

Name	w_1	w_γ
Original synthesis	1.00	0.00
Experiment 1	0.75	0.25
Experiment 2	0.50	0.50
Experiment 3	0.25	0.75
Experiment 4	0.00	1.00

Table 2. Experiments with weighting.

where I is a number of partial costs and w_i are weights with the condition $\sum_i w_i = 1$.

In our experiments we used $I = 2$, where w_1 was used for weighting the distance of a candidate unit from its cluster center, and the weight $w_2 = w_\gamma$ was used for weighting the intelligibility measure of a candidate unit. Both weights were always normalized within the cluster. In this way the selection method took into account also the intelligibility of the units, and more intelligible units should be selected from the cluster during the synthesis. The partial unit cost $d_u^\gamma(\theta_j, t_j)$ is the SII of the unit θ_j minus 1 (to get minimal value for higher intelligible units), while t_j is the reference pink noise. Fig. 1 shows this selection model in detail.

3. EXPERIMENTS

3.1. Weighting

A key issue is to define the weights for the unit cost. Having small w_γ, the influence of intelligibility measure will be very small, and having high w_γ, it might cause overall degradation of speech quality or hyperarticulation of synthesized speech. We studied here the influence of the intelligibility measure on the synthesized speech. We stepwise decreased the weight w_1, and simultaneously stepwise increased the weight w_γ. Table 2 overviews this process.

Additive noise has been added to generate five groups of test words (one for the original synthesis, and four for the experiments). We chose to use the RSG-10 [12] collection as a source of noises. We selected eight of the RSG-10 noises for use. To add noise, we used Guenter Hirsch's FaNT tool [13], using the "-m snr_8khz" option to calculate an unweighted, fullband SNR.

In this way we experimentally determined the weight for a distance of a candidate unit from its cluster center $w_1 = 0.25$, and the weight for the intelligibility measure $w_\gamma = 0.75$.

Style	Noise	Avg. WER
Plain	Babble	66.50 %
In noise	Babble	58.87 %

Table 3. Word-error-rate scores for plain speech and speech in noise at 0 dB signal-to-noise (SNR) ratio. Plain style represents an original synthesis method, and the second style 'In noise' represents a synthesis with the SII.

3.2. Listening Tests

To evaluate the relative understandability of speech in noise, we had ten people (six men and four women) listen to six examples of each speaking style, for a total of 24 sentences, which were randomly selected from the test part of the speech corpus. These sentences were arranged randomly, with the stipulation that the same sentence could not be heard twice. We added to all synthesized stimuli noise at 0 dB SNR using the FaNT tool, with the "-m snr_8khz" option. Babble noise from the RSG-10 noise collection was used (the source of this babble was 100 people speaking in a canteen, while individual voices were slightly audible). Audio stimuli were presented to listeners using the headphones. Listeners were asked to listen to the sentences maximum three times, and type in all of the words in the sentence that they could understand. These were then scored using word error rate (WER). The results are shown in Table 3.

The modified unit selection voice shows an improvement in understandability under 0 dB SNR, with a 7.63 % of the reduction in word error rate. Even the rate of 0 dB SNR with the babble noise is a high noise condition, the listeners clearly listened new words, which were in the plain synthesized speech not recognized.

4. CONCLUSIONS AND FUTURE WORK

Our experiments imply that simple modification of the unit cost in unit selection synthesis can improve understandability of speech delivered under poor channel conditions. The core of this technique is in usage of an intelligibility measure, for example speech intelligibility index, to score the suitability of the unit with respect to its intelligibility during selection from the speech corpus. Having the speech corpus rich enough,

there are many of realizations of the unit within a cluster with different intelligibility.

The advantage of this approach is that only the original speech corpus is necessary, there is no need for a special database. Hovewer, some refinements could be adopted to improve current method:

- We used a pink noise was as a representative noise for the calculation of the SII values. This decision resulted from our assumption that the stationary pink noise can serve as a first approximation of the non-stationary noises used in this research. Another extension of presented technique would be in the use of another representative noise and/or another intelligibility measure. Rhebergen and Versfeld have presented recently a modification of speech intelligibility index for prediction of speech intelligibility in non-stationary noise [14]. Using this meassure for calculation of the unit selection cost might increase the usage of presented technique in other noise conditions of a real world.

- Our calculation of the intelligibility values were done off-line, using one kind of noise which equally corrupts all one-third bands of speech. One possible extension would be to calculate these values in run-time of synthesis, using actual noise present in the environment.

5. ACKNOWLEDGEMENTS

I would like to thank Milan Rusko, who organized the listening tests, and made a valuable correction of this paper.

The work has been supported by the VEGA 2/5124/25 and by the Ministry of Education of the Slovak Republic in a frame of the state task of science and development Intelligent Speech Communication Interface.

6. REFERENCES

[1] A. W. BLACK and P. TAYLOR, "Automatically Clustering Similar Units for Unit Selection in Speech Synthesis," in *Proc. of the European Conference on Speech Communication and Technology*, Rhodos, Greece, 1997, vol. 2, pp. 601–604.

[2] H. S. VENKATAGIRI, "Segmental Intelligibility of Four Currently Used Text-to-Speech Synthesis Methods," *Journal of the Acoustical Society of America*, vol. 113, no. 4, pp. 2095–2104, 2003.

[3] S. KOESTNER, CH. POERSCHMANN, and J. WALTER, "Eine Datenbank fr deutsche Sprache mit Lombard-Effekt," in *Fortschritte der Akustik - DAGA 2000*, DEGA in Fortschritte der Akustik DAGA 2000, Ed., pp. 356–357. DEGA e.V, D - Oldenburg, 2000.

[4] B. LANGNER and A. W. BLACK, "Creating a Database of Speech in Noise for Unit Selection Speech Synthesis," in *5th ISCA Speech Synthesis Workshop*, Carnegie Mellon University, Pittsburgh, 2004, pp. 229–230.

[5] B. LANGNER and A. W. BLACK, "Improving the Understandability of Speech Synthesis by Modeling Speech in Noise," in *Proc. of ICASSP*, Philadephia, USA, 2005, vol. 1, pp. 265–268.

[6] T. DUTOIT and M. CERNAK, "TTSBOX: A Matlab Toolbox for Teaching Text-to-Speech Synthesis," in *Proc. of ICASSP*, Philadephia, USA, 2005, vol. 5, pp. 537–540.

[7] S. HAWKINS, S. HEID, J. HOUSE, and M. HUCKVALE, "Assessment of Naturalness in the ProSynth Speech Synthesis Project," in *IEE Workshop on Speech Synthesis*, London, 2000.

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

[9] E. WAN, A. NELSON, and R. PETERSON, "Speech Enhancement Assessment Resource (SpEAR) database, http://ee.ogi.edu/nsel/, Beta Release v1.0. CSLU," 1998.

[10] H. MUESCH, "SII: Speech Intelligibility Index, http://www.sii.to/, a Matlab program," 2005.

[11] X. HUANG, A. ACERO, and H. W. HON, *Spoken Language Processing : A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, Upper Saddle River, N.J., 2001.

[12] F. STEENEKEN and F. GEURSTEN, "Description of the RSG-10 noise database," Tech. Rep., TNO Institute for Perception, The Netherlands, 1988.

[13] G. HIRSCH, "Fant - Filtering and Noise Adding Tool, http://dnt.kr.hsnr.de/download.html," 2005.

[14] K. S. RHEBERGEN and N. J. VERSFELD, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2181–2192, 2005.