# 한국어 문장음성합성 시스템의 평가를 위한 다음절 무의미단어의 생성및 평가에 관한 연구

조 철 우[1o], 김 경 태[2], 이 용 주[3]
창원대학교 제어계측공학과, 한남대학교 정보통신공학과,
원광대학교 전산기공학과

# GENERATION OF MULTI-SYLLABLE NONSENSE WORDS FOR THE ASSESSMENT OF KOREAN TEXT-TO SPEECH SYSTEM

Jo, Cheol-Woo[1] Kim, Kyung-tae[2] Lee, Yong-Ju[3]
Changwon National Univ, Hannam Univ, Wonkwang Univ

ABSTRACT

In this paper we propose a method to generate a multi-syllable nonsense wordset for the purpose of synthetic speech assessment and applies the wordset to assess one commercial text-to-speech system. Some results about the experiment is suggested and it is verified that the generated nonsense wordset can be used to assess the intelligibility of the synthesizer in phoneme level or in phonemic environmental level. From the experimental results it is verified that such multi-syllable nonsense wordset can be useful for the assessment of synthesized speech.

## 1. INTRODUCTION

The end purpose of research about speech synthesis can be an implementation of system which can make a speech from given text or text-to-speech system. Many kinds of Korean text-to-speech system or rule based synthesis system are proposed , implemented and even commercialized. Some of them are considered to have reasonable intelligibility. But there are few system, whose objective test results for the intelligibility and the naturalness of their output speech were reported. Even in the case their performance is measured, the tests are almost concentrated on the measurment of intelligibility from single syllable structure.[1][2][3] There are also reports from testing foreign language systems such as in European SAM project etc. [4], [5], [6], [7], [8], [9] There are also activities to make standards of assessment in speech technology.[10], [11], [12], [13] But different language structure and constraints prevented adopting their procedure directly to our language. Testing a synthesizer is definitely not easy. That job is time consuming and even boring. The major problems that lies on the design and performing experiments are as follows. How can we make good test wordset with reasonable size? How can we gather subjects? How can its results can be helpful for later research? How small wordset can be used for testing various characteristics of synthesizer? What size of the wordset is suitable for the test not making the subjects feel boring? To decrease these problems we designed a method to generate multi-syllable nonsense wordset for testing a Korean text-to-speech synthesizer. This paper describes a method to generate multi-syllable nonsense wordset ,test procedures using the wordset and test results from the test.

## 2. DESIGN OF NONSENSE WORDSET

### 2-1. Backgrounds

In European SAM project, they used single syllable nonsense wordset to test the intelligibility of synthesizers in segmental tests. The good point of using nonsense wordset is well described on their report and they got a good result from the method.[1][2][3] When we are trying to adopt the method in Korean synthesizer, we met some problems. Some of them are peculiar to Korean language itself and some are not.

At first, total number of syllables for the nonsense wordset was too much for proper experiment.

Second, we wanted the wordset should include all the kinds of syllable structure appeared in Korean speech, but that was not possible with the algorithm suggested in SAM project. This is mainly because of the syllable structure of Korean. Korean words do not allow continuous concatenation of consonants except for the few cases. So the number of cases, that is allowed by monosyllable nonsense wordset, is very limited.

Third, we wanted the wordset be balanced to phoneme as well as to phonemic environment.

Fourth, we wanted the wordset be multi-syllable structure to test various conditions in a single wordset at a time. Current monosyllable structure cannot test various conditions simultaneously.

To meet these conditions mentioned above, we designed a new wordset suitable for our experiment.

### 2-2. Procedures to Design Wordset

To make a valuable wordset, we set three conditions.

First, they should be nonsense words.

Second, the word size should be variable for proper experiment in every case.

Third, they should be environmentally balanced as well as phonemically.

From the statistic results from Korean pronunciation dictionary, we get the probability of appearance on each phoneme and on each phonetic environment. Then 12 basic syllable structure which has similar word structure ratio as in

large vocabulary.
Phonemes are selected randomly from the list. Then substitutted to pre-selected word structure. Any multiple number of 12 wordset can be generated. Of course words are all different ones.

## 3. EXPERIMENTS AND RESULTS

Table1. shows examples of basic word structures. Using these basic word structures, 48 words are generated. The number of words is chosen for the convenience of experiment. From the Korean pronunciation dictionary appearance ratio for CVC:VCV:CC:VV is approximately 4:2:2:1.[14]

Table 1. Basic Syllable Structures

| Word Structure | CVC | VCV | CC | VV |
|---|---|---|---|---|
| CVC-VV-CVC | 2 | 2 | 0 | 1 |
| CVC-VCV-CVC | 4 | 3 | 0 | 0 |
| CVC-CVC-VV | 2 | 1 | 0 | 1 |
| CVC-CVC-VCV | 3 | 2 | 1 | 0 |
| VCV-CVC-CVC | 3 | 2 | 2 | 0 |
| VCV-VCV-CVC | 0 | 2 | 0 | 1 |
| VV-CVC-CVC | 2 | 1 | 0 | 1 |
| VV-CC-VCV | 1 | 1 | 1 | 1 |
| VV-VCV | 0 | 1 | 0 | 2 |
| CVC-CVC-CVC-CV | 3 | 0 | 3 | 0 |

In the listening experiment, 13 subjects ( 1 female, 12 male ) with normal hearing are chosen. To make them familiar with the synthetic speech sound, some synthetic speech is given them before test. Test is done in a language laboratory, each subject wearing headset. The subjects are requested to transcribe nonsense words on the paper. Two of the subject are assigned to test naturally pronounced words. Synthetic speech is recorded on Sony's DAT recorder, then suggested to subjects using facilities in the laboratory. Before the test, each booth is tested for the proper sound levels and levels are adjusted. In the experiment, subject are acknowledged to know that the words are all nonsense words to reduce any possible wrong results. Korean commercial text-to-speech synthesizer "GARASADAE" is used to be assessed
From the experiment, each phoneme's recognition ratio is between 5.2 and 75.0% in synthetic speech and between 24.1 and 87.5% in natural speech. And the recognition ratio is high in the sequence of VV, VCV, CC, CVC in synthetic speech by decreasing order while it is VV, VCV, CVC, CC in natural speech. So we can conclude that the suggested wordset have similar characteristics in the aspect of recognition by subjects. But as we can see from the table 2, some kinds of environments are very unstable in its recognition ratio.

For example subject C have much lower recognition ratio, we can conclude from this fact that this subject have a stronger trend to give his or her own meaningful word despite of previous precautions. So these data are better be removed from final statistics. When comparing natural and synthetic speech recognition ration between these two groups are 38.4%, 38.27%, 28.6%, 26.9% in VV, CC, VCV, CVC respectively. CVC shows the lowest recognition ratio, so we can conclude that such phonetic environment should be analyzed more precisely. Also relative recognition ratios for each phonemes are shown on figure 1.

Table 2. Environmental Recognition Ratio (%)
<Synthetic Speech>

| Speaker | CVC | VCV | CC | VV |
|---|---|---|---|---|
| A | 4.9 | 12.3 | 16 | 31 |
| B | 8 | 18.5 | 15.5 | 31.3 |
| C | 6.6 | 6.2 | 5.2 | 12.5 |
| D | 12.4 | 16.7 | 12.1 | 43.8 |
| E | 14.1 | 16.7 | 10.3 | 18.8 |
| F | 6.6 | 15.2 | 1.7 | 18.8 |
| G | 6.6 | 12.1 | 10.3 | 12.5 |
| H | 18.2 | 25.8 | 13.8 | 75 |
| I | 13.2 | 25.8 | 8.6 | 50 |
| J | 10.7 | 18.2 | 12.1 | 43.8 |
| K | 10.7 | 18.1 | 12 | 43.7 |
| Average | 10 | 15.8 | 10.2 | 31.2 |

< Natural Speech>

| Speaker | CVC | VCV | CC | VV |
|---|---|---|---|---|
| NA | 32.2 | 43.1 | 24.1 | 75 |
| NB | 42.2 | 63.6 | 29.3 | 87.5 |
| Average | 37.2 | 53.36 | 26.73 | 81.25 |

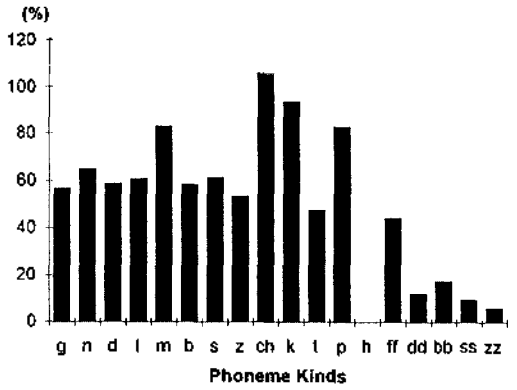


Fig. 1 Relative Recognition Ratio (%)
(Synthetic/Natural)

Generally relative recognition ratio for the phonemes /g/,/n/,/d/,/r/,/b/,/s/,/z/are between 36% and 58%. But for /ch/ synthetic speech shows higer recognition ratio. This is unexpected result normally. But this can be explained as follows. Generally in synthesizer plosives or fricative are strongly emphasized in unvoiced portions than in natural speech. So this can be a quite natural result. Also strong consonants such as /gg/, /dd/, /bb/, /ss/, /zz/ shows very low recognition ratio between 1.1% and 9.1%. This result shows

us such kind of phonemes are very difficult to synthesize in the synthesizer we used. So much of the efforts should be given to analyze these phonemes in later works. But natural speech is recognized relatively highly for the same kinds of consonants. Table3 and table4 is a confusion matrix for natural and synthetic phonemes respectively. In table3 recognition ratio for the strong consonants is very low. This can be caused from the reason for the same consonants have a lower recognition ratio. Phoneme /h/ is not recognized at all in synthetic speech. This may be caused by the reduction rule of phonemes between specific phonemes in Korean. Also this may be caused by not considering to include this kind of inter-phonemic reactions in wordset generation.

The nonsense wordset used in this paper has a similar appearance ratio in phonemic environments as a large size word dictionary. Also appearance ratio for each phonemes are similar even though this is not an intentional one. So this wordset can be useful to assess the synthetic speech considering phoneme environments. And test results shows a reasonable result phonetically and this will be useful for the assessment of individual intelligibility. Also in the comparison between synthetic and natural speech, the result shows a similar sequence. In conclusion this wordset can be useful for the assessment of synthetic speech.

But there are still some points to be improved and further developed as follows.

(1) When generating wordset. General phonemic concatenation rule should be applied. Namely the appearance probability, such that a phoneme comes before or after a phoneme, should be considered. At this stage we considered such an effect by manual correction after automatic generation.

(2) Creation or variation of new phonemes. When a phoneme is put to some phonetic environment, should be considered.

(3) Some minor phonetic environments that are not considered in this paper should be included.

(4) Reasonable mixture of synthetic and natural speech in the same list of words in an experiment.

(5) Equalize appearance ratio of each phoneme.

(6) Reduction or minimization of prosodic effects inside the word.

In general, 4 syllable nonsense words have nearly the same prosody in synthetic speech. In this experiment, the synthesizer showed the same decreasing trend of pitch on every nonsense word. But designing a wordset to meet all these conditions is a very complex problem and these should be paralleled continuously by the phoneticians, phonetic engineers etc. So limiting test condition like in this experiment will be prerequisite at this stage.

## 4. CONCLUSIONS

In this paper we suggested a method to generate a multi-syllable nonsense wordset to assess Korean synthetic speech and showed the results.

At first we limited conditions that the wordset should have, and generated a nonsense wordset by referencing a statistic results on the Korean phonemes. Total 48, 4 syllable nonsense wordset is generated. And the speech for each word is recorded and suggested to 13 subjects. Listening test is carried out in university language laboratory. From analysing the result out of listening test, we can conclude that the suggested wordset have the similar characteristics with natural wordset, also similar appearance for each phoneme. Also the results represent phonetical correctness.

Concludingly suggested method can be useful to generate nonsense words and to assess the synthesizer. We will improve the system and are developing a software to generate wordset and to do the test automatically.

## 5. REFERENCES

1. Kim, J.H., Jang, D.Y, Kang, S.H, "Development of Monosyllable Lists and Evaluation for Assessment of Intelligibility", Proceedings of ASK, Vol.12, No.1, 1993

2. Kang,C.H. et.al.,"Evaluation of Synthetic Speech According to Kinds of Syllables", Proceedings of ASK, Vol.12, No.1, 1993

3. Cheong,Y.H., Choi,J.H., Han,M.S., "Evaluation of Synthetic Speech from TTS", Proceedings of ASK, Vol.12, No.1, 1993

4. Multi-lingual Speech Input/Output Assessment, Methodology and Standardisation, Final Report, SAM-UCL-G004, 1992

5. Speech Input/Output Assessment and Speech Database, Proceedings of ESCA workshop, 1989

6. Louis C.W. Pols,"Evaluating the Performance of Speech Technology Systems", Institute of Phonetic Sciences, University of Amsterdam, Proceedings 15, 27-41, 1991

7. Nobuhiko Kitawaki, "Speech Quality Assessment for Speech-coding and Speech Synthesis Systems", Journal of IEICE, No.70, No 4, 1987

8. Hideki Kasuya et.al., "Factors Involved in Evaluation of Techniques for Speech Processing", Research Report No. PASL 63-8-1 supported by Grant-in-aid for scientific research on Priority Areas "Advanced Man-Machine Interface Through Spoken Language", Japan, 1988

9. Hideki Kasuya, "Assessment of Speech Synthesis Technology", Research Report No. PASL 62-8-1 supported by Grant-in-aid for scientific Research on Priority Areas "Advanced Man-Machine Interface Through Spoken Language",Japan, 1988

10. "Experiment in assessing the quality of synthetic speech", Temporary Document No.70-E, CCITT working party 5/XII, Geneva, 27 Feb.-3 March 1992

11. "Elements for Draft Rec. on synthetic speech assessment", Temporary document 80-E, CCITT working party 5/XII, Geneva, 27 Feb - 3 March 1992

12. "Report on Question 5/XII, Speech synthesis/recognition systems", Temporary Document No.52(Rev.1)-E, CCITT working party 5/XII, Geneva, 11-19 May 1993

13. "Draft Recommendation P.8S-Subjective Performance Assessment of the Quality of Speech Voice Output Devices",COM 12-6-E, CCITT working party 5/XII, 1993

14. "Korean Pronunciation Dictionary", KBS Korean Research Association, 1992

Table 3. Confusion Matrix for Natural Speech

| | g | n | d | l | m | b | s | z | ch | k | t | p | h | gg | dd | bb | ss | zz | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| g | 10 | | | | | | | | | | | | | 6 | | | | | 16 |
| n | | 5 | | | | | | | | | | | | | | | | | 5 |
| d | | | 9 | | | | | | | | | | | | | | | | 9 |
| l | | 2 | | 18 | | | | | | | | | | | 1 | | | | 21 |
| m | | | | | 18 | | | | | | | | | | | | | | 18 |
| b | | | | | | 9 | | | | | | 3 | | | | 11 | | | 23 |
| s | | | | | | | 20 | 1 | | | | | | | | | 41 | | 62 |
| z | | | | | | | | 13 | 4 | | | | | | | | | 2 | 19 |
| ch | | | | | | | | | 16 | | | | | | | | | | 16 |
| k | | | | | | | | | | 17 | | | 1 | 2 | | | | | 20 |
| t | | | | | | | | | | | 24 | | 2 | 1 | | | | | 27 |
| p | | | | | | | | | | | | 7 | | | | | | | 7 |
| h | | | | | | | | | | 1 | | | 12 | | | | | | 13 |
| gg | 5 | | | | | | | | | | | | 1 | 5 | | | | | 11 |
| dd | | | | | | | | | | | 2 | | | | 6 | | | | 8 |
| bb | | | | | | 3 | | | | | | | | | | 7 | | | 10 |
| ss | | | | | | | 2 | | | | | | | | | | 5 | | 7 |
| zz | | | | | | | | | 2 | | | | | | 1 | | | 12 | 15 |
| blank | 1 | 1 | 1 | 2 | | | | | | | | | | | | | | | 5 |
| total | 16 | 8 | 10 | 20 | 18 | 12 | 22 | 14 | 22 | 18 | 26 | 10 | 16 | 14 | 8 | 18 | 46 | 14 | 312 |

Table 4. Confusion Matrix for Synthetic Speech

| | g | n | d | l | m | b | s | z | ch | k | t | p | h | gg | dd | bb | ss | zz | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| g | 23 | | | 1 | | 2 | | 6 | | 2 | | | 3 | 22 | 3 | 5 | | 6 | 73 |
| n | 1 | 13 | 3 | 18 | 6 | 1 | | | | | | | 1 | | | | | | 43 |
| d | | | 21 | 4 | | 2 | | | | | 3 | | 1 | 5 | 17 | 3 | | | 56 |
| l | 6 | 1 | 3 | 44 | 1 | | | 1 | | | | | 2 | | | 1 | | | 59 |
| m | 1 | 5 | | 2 | 60 | 8 | | | | | | | 5 | | | 7 | | | 88 |
| b | | | | 2 | | 21 | 1 | | | | 1 | 1 | 5 | 1 | 3 | 28 | 4 | 1 | 68 |
| s | | | 2 | | | | 49 | 5 | 16 | 2 | 2 | 1 | 8 | 2 | | | 66 | 2 | 155 |
| z | 7 | | 3 | | | 1 | 23 | 28 | | | 3 | 1 | 10 | 5 | 2 | | 75 | 35 | 193 |
| ch | 2 | | | | | | 6 | 6 | 68 | | 2 | | 16 | | | 2 | 15 | 3 | 120 |
| k | 14 | | | | | 1 | | | | 63 | 27 | 7 | 1 | 6 | | 2 | | 1 | 122 |
| t | | | 6 | | | 2 | | | | | 46 | 6 | | | | 3 | 2 | | 65 |
| p | | | 1 | | | 2 | | 1 | | 1 | 13 | 23 | | 2 | 1 | 1 | | | 45 |
| h | | | | | | | | 1 | | 2 | 4 | | | | | | 2 | 1 | 10 |
| gg | 4 | | | | 1 | 1 | | | | 1 | | | | 9 | 2 | 4 | | 1 | 23 |
| dd | | | | | | | | | | | | | | | 3 | | | | 3 |
| bb | | | | | | 2 | | | | | | | | | 1 | 5 | 1 | | 9 |
| ss | | | | | | | 2 | 2 | | | | | | | | | 2 | | 6 |
| zz | 1 | | | | | | 1 | 2 | | | | | | 2 | | | 10 | 3 | 19 |
| blank | 5 | 13 | 1 | 9 | 4 | 5 | 6 | 4 | 4 | 1 | 3 | 1 | 12 | 2 | | 11 | 7 | 3 | 91 |
| total | 64 | 32 | 40 | 80 | 72 | 48 | 88 | 56 | 88 | 72 | 104 | 40 | 64 | 56 | 32 | 72 | 184 | 56 | 1248 |