Speech Rate Control for Improving Elderly Speech Recognition of Smart Devices

Guiyoung SON[1], Soonil KWON[1,*], Yoonseob LIM[2]
[1]*Dept. of Digital contents, Interaction Technology Lab., Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Korea*
[2]*Center for Robotics Reseasrch, Robotics and Media Institute, Korea Institute of Science and Technology, Seoul 02792, Korea*
[*]*skwon@sejong.edu*

***Abstract*—Although smart devices have become a widely-adopted tool for communication in modern society, it still requires a steep learning curve among the elderly. By introducing a voice-based interface for smart devices using voice recognition technology, smart devices can become more user-friendly and useful to the elderly. However, the voice recognition technology used in current devices is attuned to the voice patterns of the young. Therefore, speech recognition falters when an elderly user speaks into the device. This paper has identified that the elderly's improper speech rate by each syllable contributes to the failure in the voice recognition system. Thus, upon modifying the speech rate by each syllable, the voice recognition rate saw an increase of 12.3%. This paper demonstrates that by simply modifying the speech rate by each syllable, which is one of the factors that causes errors in voice recognition, the recognition rate can be substantially increased. Such improvements in voice recognition technology can make it easier for the elderly to operate smart devices that will allow them to be more socially connected in a mobile world and access information at their fingertips. It may also be helpful in bridging the communication divide between generations.**

***Index Terms*—automatic speech recognition, human computer interaction, speech analysis, man machine systems, human factor.**

I. INTRODUCTION

In today's modern society, advancements made in medicine have led to a growing aging population. Some recent studies on the aging population show that the elderly are beset by loneliness, a sense of exclusion and a lack of assistance in leading their daily lives [1]. As the social roles of the elderly becomes disconnected with society, there have been calls for establishing a better support system for senior citizens and launching planned welfare programs for them. However, the rise of new media has largely left the elderly behind. This has, in effect, worsened the social disconnect that the elderly are experiencing.

Smart devices are widely available and accessible in many respects. The many features installed on these smart devices allow people to manage their schedule, use financial services, play games, and do many more productive tasks that are relevant to everyday life. The young find it easy to learn and adopt new technologies introduced in smart devices that enhance their social life, but that is not the case for the elderly, who find new technologies difficult to grasp, and thus, leaving them in a difficult position to stay socially connected. The lack of operational manuals and instructions in many of these devices has compounded the problem that the elderly face in narrowing the digital divide [2]. In terms of the input method in current devices, touch-based interfaces are just as difficult to use for elderly users as button interfaces.

Voice recognition is the most important technology that the elderly need when using a smart device. Voice recognition lets a user operate a device in real-time through voice commands. It requires only a simple knowledge of how voice recognition works for elderly users to effectively begin using smart devices with relative ease. Despite the fact that most current smart devices come with voice command interfaces powered by voice recognition, the preinstalled voice recognition system has difficulty in correctly recognition the voices of the elderly, leading to low recognition rates. The reason for this is that the voice recognition system is generally optimized to recognize the speech of the young, and thus, struggles to recognize the speech style of the elderly. This research study will measure the changes in the recognition rate before and after the speech rate by syllables is modified and then analyze the results to identify the reasons for the changes.

This composition of this paper is as follows. In Chapter 2, previous studies on the aging process of vocal organs among the elderly are summarized, and then Chapter 3 introduces the methods on determining the boundaries of each syllable. Chapter 4 explains Synchronized Overlap-Add Algorithm (SOLA), a method that allows for voice signal modification of speech rate without distortion. An analysis of the before-and-after elderly voice recognition performance results derived from speech rate modification by syllable is presented in Chapter 5. Lastly, Chapter 6 details the future direction of research and includes the conclusion of this research paper.

II. EXISTING RESEARCH

The purpose of this paper is research on voice signals in the speech of the elderly based on previous engineering research related to voice recognition as well as non-engineering research in such fields as speech linguistics and biomechanics. As humans age, their lung functions and muscles in the occipital region become weaker in addition to the thinning of the mucous membrane along the vocal cords that undergo keratinization. Another aging process is the ossification of the thyroid cartilage among men aged over 65. With women over 65, the ossification process is limited to the lower part of the thyroid [3]. In elderly people, the

This research work was partly supported by the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korean government (MSIP) (No. R0126-15-1119, Development of a solution for situation awareness based on the analysis of speech and environmental sounds)

Digital Object Identifier 10.4316/AECE.2017.02011

thickness of the tongue and the duration of movement is reduced due to aging [4]. As a result, the elderly slow speech rate, silence lengths and have decreased speech accuracy [5]. In this case, due to such ossification, the speech rate of the elderly who are over 65 displays a considerable drop in intonation speed compared to that of the young. The aging process also changes the resonance characteristics of the larynx, which leads to a more thin and raspy voice [5]. All these factors combined together increase fluency breaks as well as higher frequency and length of inter-syllabic silence [6].

Ryan [7] demonstrated that a group of elderly people aged between 70 and 80 displays an overall reduction in reading speed and speech rate in addition to an increase in average voice intensity compared to that of a the young of adults. Kim et al found significant decreases in reading speed between a group of elderly people aged between 60 and 70 and a group of people aged between 40 and 50 [8]. Also, fluency breaks increase, and silence lengths are longer and more frequent [9].

According to the experiment conducted by Harnsberger [10], it is possible to show the acoustic pattern of voice by age in a comprehensive way. The main factors that he cited were fundamental frequency (F0) and speech rate. In other words, he concluded, people can deduce the age of a person by analyzing these two factors in that person's voice [8].

Based on the abovementioned existing research, it is possible to predict that there will be a difference between the voices of the young and that of the elderly due to numerous factors such as speech rate, silence length, fluency breaks and vocal frequencies. A few of these factors have already been acknowledged to reduce the accuracy of voice recognition technology. People who talk fast can have their voices modulated by Cepstrum Normalization to prevent performance reductions in voice recognition technology [10]. By modeling the changes in speech rate, the performance of voice recognition technology, which has had difficulty with speech rates, can be increased by 1.9% [11, 12].

The elements mentioned above have not been comprehensively researched as of yet. There is especially a lack of detailed proof of any connection made between voice recognition and the human voice undergoing the aging process. We recently carried out an experiment to analyze the relation and performance comparisons of smartphone-based voice recognition technology and the speech rates that have been affected by the aging process [13]. The average speech rate of the young in their 20's was used as the baseline reading for modifying the speech rate of the elderly in the experiment. After the speech rates of the elderly were modified, the success rate in voice recognition of elderly males increased to 77.9% from 76%, showing a 1.9% increase. For elderly females, the success rate rose to 76.4% from 75.4% for a 1.0% increase. In the previous experiment results, the success rate did not improve significantly despite increasing the speech rate of the elderly group. The reason for this dismal improvement was that the speech rate modification was applied across the voice data uniformly using an average length. The average speech rate was derived by dividing the length and number of syllables. Hence, the average rate could show a large gap between the speech rates by each syllable. Therefore, boosting the overall speech rate will allow slowly-pronounced syllables to be recognized more accurately while syllables originally spoken at a faster speed will be recognized for a different word or fail to be recognized entirely.

III. METHOD

To automatically segment speech signal into distinct syllables, we first calculated the probability of syllable transition that can be used to classify the syllable transition boundaries for each frame and then chose the local peaks of the smoothed traces of syllable transition probability as syllable segmentation points.

A. Syllable transition probability

We constructed a convolutional neural network (CNN) to generate feature vectors that are fed into a fully-connected network (FC) for frame-by-frame syllable transition boundary classification (Figure 1a). For normalization within CNN, two different max-pooling layers are included. Input to CNN is the Gammatone Frequency Cepstral Coefficients of each frame of sound (GFCCs, the number of channels: 128, center frequency range: 50 Hz-4 kHz, 32 coefficients) [14]. The width of input feature is fixed at 32, which corresponds to 32msec in time. Each GFCC is tagged with classification label, 0 or 1 (0 means that current frame contains no annotated segmentation point or vice versa).

To train the network, we first divided the dataset into training and test sets for different age group. Training and test set contain voices from different people (Training set: 32 speakers, Test set: 8 speakers). We did not differentiate the sex of speakers. Within the training set, we performed 5-fold cross validation to build the best model for syllable segmentation. In this paper, we only report the result with the test set data (Table I). To minimize the overfitting during training, the dropout rate of fully connected network is chosen as 0.5, and whole network is trained using stochastic gradient descent method through backpropagation Probability of syllable transition is calculated by Softmax at the final layer of fully-connected network (example trace is shown on green curve in Figure 1b).

B. Syllable segmentation

Output of the trained network contains the classification result of syllable transitions for each frame of sound. To select single point among several possible transition points as a syllable segmentation point, we smoothed the syllable transition probability using a sliding rectangular window (70 msec) and chose the local maxima (minimum peak distance: 50 msec). To quantify the segmentation performances, we are defined three different measures: detection rate, insertion rate, and missing rate. Detection and missing rate correspond to the ratio of detected or missed boundaries among manually annotated boundaries. Insertion rate is defined as the ratio of peaks of syllable transition probability where no boundaries are assigned. To determine whether or not selected boundary can be assigned as segmentation boundary, the threshold of temporal deviation window is also defined and 100 msec was used in this study. Temporal scale for smoothing and boundary deviation threshold used in this study were determined based on the overall performance of these three measures.

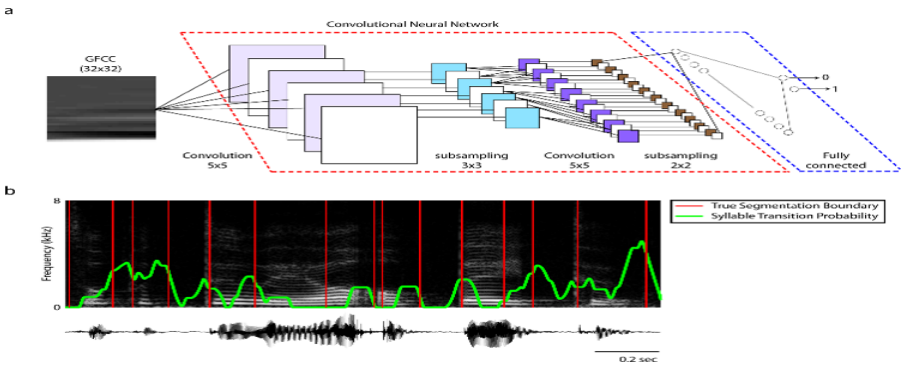

Figure 1. Procedure to obtain syllable segmentation of sound. (a) Convolutional neural network is trained to classify each frame of sound as syllable transition point or not. Input to neural network is GFCC of sound with fixed size (32x32). (b) Using trained neural network, syllable transition probability is calculated for each frame of sound. We chose peaks of the probability trace as syllable segmentation points.

IV. Speech Rate Modification

The Synchronized Overlap-Add (SOLA) algorithm is a basic method used to adjust the speech rate according to measured ratio on a time-scale[13,15,16]. It uses pitch information in order to overlap two neighboring windows that contain a series of data that is added, after which it converts the speech rate without changing the characteristics of the vocal frequency. In Figure. 2 the series of data length winlength used to run through the SOLA calculations is shown. ***Sa*** denotes Analysis Shift, which is the segmentation length when the signal analysis begins; Kmax is the maximum movement limiter for searching the pitch and is used to match the two pitch wavelengths in between each window. The change in the speed rate is expressed as a=***Ss/Sa***. If "a" is less than 1 (a<1) then the speech rate is faster than the source speed. If "a" is greater than 1 (a>1), the speech rate is slower than the source speed. The change in the speed rate or "a" is limited between 0.5 ~ 2.0 [17-19].

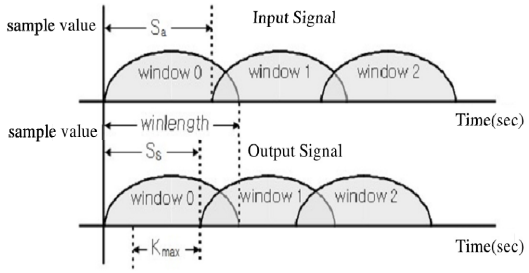

Figure 2. Example of modification method of speech rate (SOLA)

V. Experiments and Results

For the purpose of data analysis, 20 male and 20 female adults aged between 20 and 30 who had no speech disabilities, nerve disorders, vocal cord conditions, or impaired vision were selected as subjects. In the elderly group, the subjects were 20 male and 20 female adults over the age of 65. The entire group came to 80 subjects. The voice recordings for the experiment were conducted in a quiet room at the Gwangjin Senior Citizen Welfare Hall, Seoul. The microphones were kept at a distance of 20cm from the subjects. The words chosen for the recordings were the 50 most commonly-used features found in the manual of a smart device manufactured by company S. The word selection included 12 words with two syllables, 15 words with three syllables, 11 words with four syllables, 4 words with five syllables, 5 words with six syllables, 1 word with seven syllables, and 2 words with eight syllables that rounded out the 50 words. The recorded voice files were then tested with a smart device that had built-in voice recognition capabilities. When the voice recognition generated a word that correctly matched the recorded word, this was considered a successful voice recognition. Any other cases that deviated from the recorded word were judged as an erroneous recognition rate.

The proposed method was implemented to preprocess speech signal coming into the microphone installed on smart devices followed by speech recognition. The preprocessing procedure as follows:

1. Segment speech signal into syllables.
2. Calculate the length of each segment.
3. Modify the length of each segment with respect to the normal length.

A. Syllable transition boundary classification performance

Although syllable segmentation point is determined based on only the acoustical structure of signal, GFCC, we found that the proposed method shows a moderate level of performance in classifying each frame of the sound signal that contains syllable segmentation point or not. Table I shows the frame-by-frame performance of the proposed classification method for recordings of test set. Positive in this table means that input contains syllable transition point and vice versa. The average accuracy of proposed method is around 71.6%. Elderly people often prolong the vocalization in time which can create changes in spectrotemporal structures of sound. However, using only GFCC feature, proposed system also shows the similar performances for different age groups (note that average accuracy for elderly people voice is 71%).

TABLE I OVERALL PERFORMANCE OF THE TRAINED NEURAL NETWORK FOR SYLLABLE TRANSITION BOUNDARY CLASSIFICATION.

	The young voice		The elderly voice	
	Classified +	Classified -	Classified +	Classified -
Actual +	70.7%	29.3%	70.1%	29.9%
Actual -	26.4%	73.6%	28%	72%

B. Syllable segmentation performance

Frame-by-frame classification of syllable transition point can classify several consecutive frames as syllable segmentation point. To reduce such multiple assignments for syllable segmentation, we smoothed syllable transition probability traces using rectangular window (width: 70 msec) and pick the local maxima as final syllable segmentation points. In Figure 3, we have illustrated examples of segmented voices of the young and elderly people chosen from the test set. Green curves correspond to the smoothed traces of syllable transition probability. Blue solid (or dotted) lines represent the segmentation boundaries that are assigned as detected or inserted. Red solid (or dotted) lines represent the segmentation boundaries that are assigned as annotated or missed. Overall, smoothed syllable probability trace for the young voice corresponds well with the manually annotated boundaries. When we cross-correlated syllable transition probability with annotated boundary for the young voice (smoothed by the same window), average value of maximum cross-correlation ratios was 0.72. However, for elderly people voice, mean cross-correlation was only 0.59. This indicates that proposed method could have more segmentation boundaries that may be assigned as being missed or inserted for elderly people voice. Also, the method fails to detect segmentation points where off-set and on-set of syllables are closely neighboring each other (these points are indicated by arrows in Figure 3).

To quantify the segmentation performance, we defined three different types of segmentation boundary as detected, missed, or inserted based on the temporal distance [20]. For example, when the time difference between selected boundary and the closest true boundary is within the temporal deviation threshold, corresponding true boundary is tagged as detected. Figure 4 shows the distribution of temporal deviations between the selected boundary and the closest manually annotated boundary.

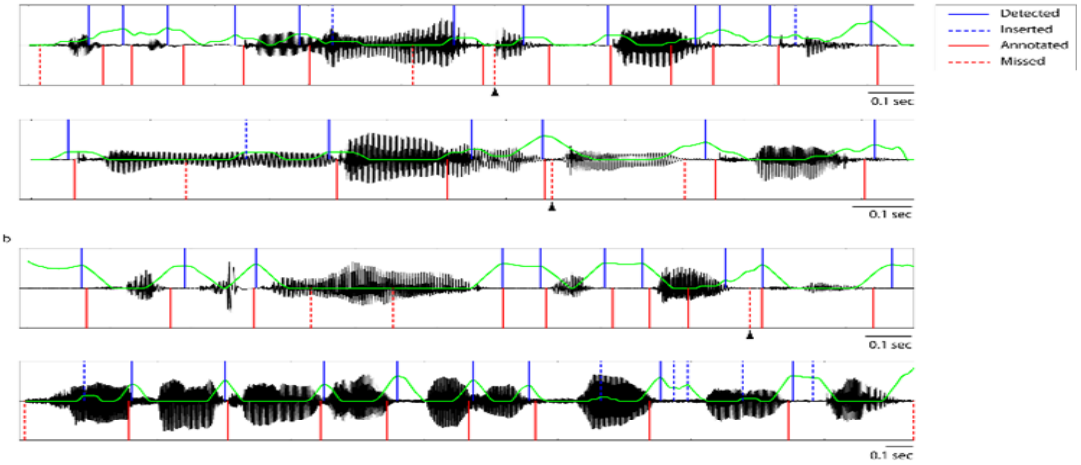

Figure 3. Syllable segmentation results. (a) Segmentation results of the young voice. (b) Segmentation results of elderly people voice. Blue lines represent detected boundaries by proposed method, and red lines represent manually selected syllable boundaries. Dotted lines for each color represent inserted (blue) or missing (red) boundaries. Green curve is the smoothed trace of syllable transition probability

More than half of the selected boundaries exists within 30 msec from true segmentation point. From these measures, we found that proposed method can detect 85.3% of manually annotated boundaries for the young voice (79.5% for elderly people voice, Table II). However, we found that more than 20% of local maxima are selected as inserted the boundaries and segmentation performance for elderly voice is lower than that for the young voice.

TABLE II OVERALL PERFORMANCE OF THE PROPOSED METHOD FOR SYLLABLE SEGMENTATION

	Detection rate	Missing rate	Insertion rate
The young	85.3%	14.7%	23.9%
The elderly	79.5%	20.5%	28.5%

To confirm the differences in the success rate of voice recognition of the elderly after speech modification was applied to each syllable, the boundaries in the placement of syllables from spoken words by the elderly and the young was collected. First, a statistical analysis of the data was carried out to derive the average length of the syllables spoken by the young. The average rate of speed among the young was used as the baseline reading to which the SOLA was applied to modify the speech rate of each syllable spoken by the elderly group to mimic that of the young's rate. The graph in Figure 5 shows how much the success rate of voice recognition increased after speech modification to each syllable in the elderly group was made to follow the speech rate patterns of the young closely.

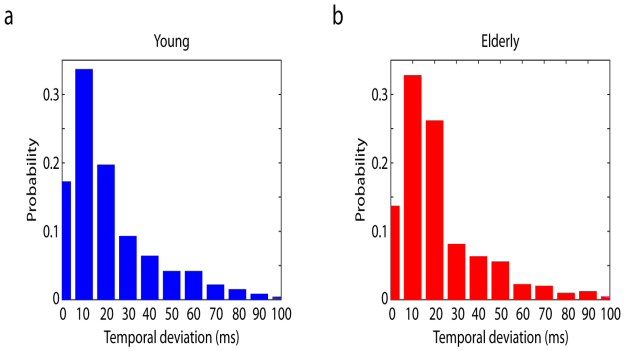

Figure 4. Distribution of temporal deviation of segmentation point.
(a) Temporal deviation of syllable segmentation for the young voice.
(b) Temporal deviation of syllable segmentation for the elderly voice

Based on the experiment results, the voice recognition success rate for elderly males was 78.6% while elderly females recorded a success rate of 77.0% before speech modification was applied to each syllable. When compared to the recognition success rate of voice data derived from the young, there was an approximately 17% difference in the recognition success rate. However, after speech

modification was applied to each syllable, the success rate jumped to 93.1% for the elderly males and 87.0% for that of the elderly females. Therefore, the elderly males and elderly woman recorded a 14.5% and 10.0% increase respectively after speech modification was applied to each syllable. In addition, when comparing these figures to the success rate of voice data derived from the young, there was only a 1.2% difference between the elderly males and the young while the gap narrowed to 7.9% between elderly the females and the young.

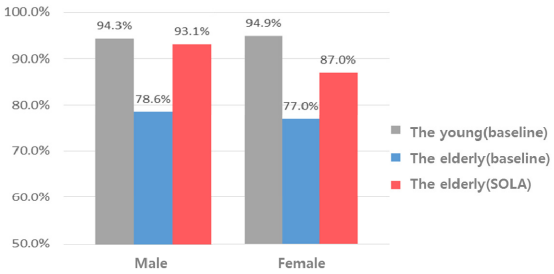

Figure 5. Voice recognition success rate of the young and voice recognition success rate of the elderly before and after speech medication of each syllable

We calculated the average speech rate of the young and the elderly resulting in 5.26 syllables per second and 3.44 syllables per second. In addition, as we increased the rate of the elderly speech by 5, 10, 20%, corresponding to 3.61, 3.79, 4.13 syllables per second, the speech recognition accuracy peaked at 10% (Figure 6): 3.79 syllables per second is 72% of the speech rate of the young used in our experiments. It means that there is a limit to improving speech recognition accuracy by simply making the rate of the elderly speech faster.

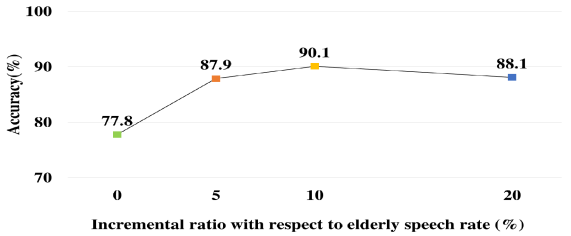

Figure 6. Voice recognition of incremental ratio with respect to the elderly speech rate

Even after applying speech modification to each syllable for the elderly including both male and female subjects, the average rate of failure for the entire elderly remained at 9.9%. After analyzing the reasons for this failure in voice recognition, weakened initial consonant sounds and the improper inter-syllabic silence was identified as the two key contributing factors. Mispronounced words also caused failure in voice recognition and is a completely separate issue from speech modification processing.

In Figure 7, the "Weakened initial consonant" denotes cases where the voice signal of the initial consonant became weak when the syllable of the initial consonant was shortened due to the speech modification process of each syllable. For example, the consonant letter, "M" in the compound word, "Mi Le Khol Sel Ceng" (Mirror call setting) was not pronounced properly and registered as "I Le Khol Sel Ceng." In the case of the word "Nal Cca," (Date) the double "C" was not pronounced correctly and thus, failed to be recognized. After looking at the voice recognition results for the word, "Nal Cca," the word registered as "Nal Ca"(Fly), which the voice recognition system made an approximate guess to the closest word to it, which was "Nam Ca"(Man). Among the recognition errors, after speech modification was applied to each syllable, "Weakened initial consonant" took a 23.6% share of all errors. To resolve this problem, the syllables of initial consonants needs to be managed separately with a focus on retaining its syllable length and energy.

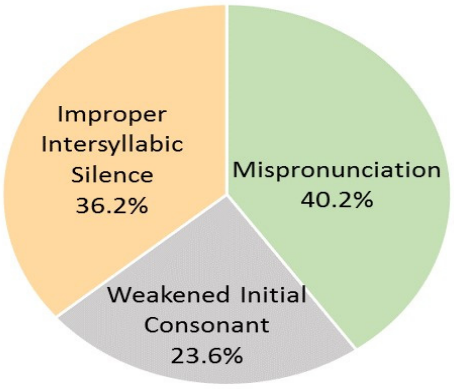

Figure 7. Contribution factors to errors in voice recognition after speech modifications by each syllable

The "Improper inter-syllabic silence" in Figure 7 is a case where the intersyllabic silence was improperly placed and does not include prolonged sounds where there should be no silence. Based on the data analysis used in this research, improper intersyllabic silence should last for over 0.02s on average. "Improper intersyllabic silence" occurs when the silence length between syllables is too short or completely absent after speech modifications were made to each syllable. These cases comprised 36.2% of all failed voice recognition rates. In order to resolve this issue, the SOLA was modified so that a minimal inter-syllabic silence could be maintained. In Figure 7, "Mispronunciation" denotes cases that the experiment subject mispronounced the given word. One example was the mispronunciation of "I Pu Ssu Li Mo Pa Il," which incorrectly pronounced as "Pu I Ssu Li Mo Pa Il" (V Three Mobile). These mispronunciation cases comprised 40.2% of the recognition failure data after speech modification was applied to each syllable and thus was the biggest contributor.

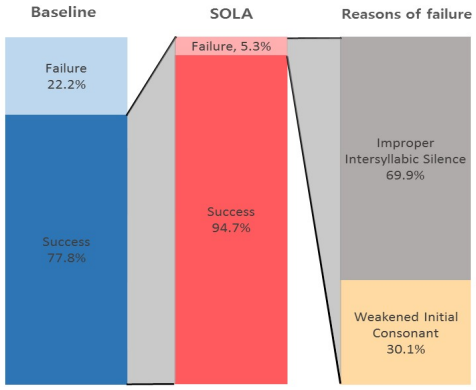

Figure 8. Analysis of failure of voice recognition after speech modification and success of voice recognition when no speech modification was applied to each syllable

Figure 8 shows an analysis of how the voice recognition failed after the speech rate was modified despite succeeding when no modification to the speech rate was applied. On the other hand, Figure 8 shows the reason for the failure in voice recognition experiment after the speech rate was

modified as well as before it was modified. In contrast to the example in Figure 9, Figure 8 shows that the proposed method did not fail to recognize the speech due to any "mispronunciation." If there were any mispronunciations, then the voice recognition experiment would have failed before any modifications were applied to the speech rate.

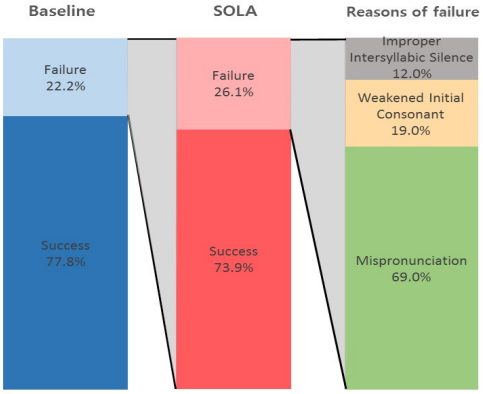

Figure 9. Analysis of failure in voice recognition after speech modification as well as before speech modification was applied to each syllable.

In the case of the "Improper inter-syllabic silence" issue, it became more pronounced when voice recognition succeeded with a speech rate before the speech rate being modified by each syllable instead of when it failed before the speech modifications. The words that were successfully recognized in the voice data before speech modifications were made by each syllable were compounds nouns such as "Ta Wun Lo Tu Kwan Li"(Download management), "Mi Le Khol Sel Ceng"(Mirror call setting) and "So Phu Thu Wey E Ep Tey I Thu"(Software update). Such words comprised 72.4% of words with a successful recognition rate. Such compound nouns are naturally spoken with a brief silence between the two nouns and the voice recognition system is optimized for this. When there is no silence between the two nouns, the voice recognition system will fail to recognize the word. To prove this phenomenon further, the set of compound words that failed to be recognized were manually inserted with silence in between the nouns and retested for voice recognition, and there was an 84% success rate in voice recognition among words classified for having "Improper inter-syllabic silence" .

VI. Conclusion

The purpose of this research paper is to enhance the usability of smart devices among the elderly by correcting shortcomings in the voice interfaces used in smart devices. The paper also presents the experiment results. There are a multitude of factors that lower the rate of success in voice recognition, however, in the case of the elderly, a key contributing factor is their improper and unnatural speech rate. To correct this problem, modifications to the elderly's speech rate was used in the preprocessing for voice recognition and resulted in an average increase of 12.3% in the successful recognition rate.

Despite showing meaningful results in this research, there is one area that we would rather have liked to make progress in: the need for a more accurate method for syllable segmentation for real world application. Once this is resolved, the current voice interfaces that are optimized for the young will become easier to use for the elderly. In addition, resolving this issue will increase the usability of smart devices among the elderly and thereby contribute to narrowing the generational gap between the young and elderly while helping the elderly become more socially connected.

Acknowledgment

Thank you for anonymous reviewers.

References

[1] Korea National Statistic office. "Social Survey; Welfare Category; Difficulties Experienced by Senior Citizens, Official Statistics Research Newsletter, vol.5, pp.2-3, 2013.

[2] W. S. Kang, M.S. Kim, J.W. Ko, "Effects of the smartphone information use and performance on life satisfaction among the elderly," Korean Gerontological Society, vol.33, no.1, pp.199-214, 2013.

[3] B.C. Sonies, "Oral-motor Problems," Communication Disorders in Aging: Assessment and Management, Washington, Gallaudet University Press, pp. 185-213, 1987.

[4] J.W. Bennett, P.H.H.M. Van Lieshout, C.M. Steele, "Tongue control for speech and swallowing in healthy younger and older subjects," International Journal of Orofacial Myology," vol.33, pp.5–18, 2007.

[5] J.C. Kahane, "Anatomic and physiologic changes in the aging peripheral speech mechanism," Aging: Communication processes and disorders, pp.21-45, 1981.

[6] S. Y. Lee, "The overall speaking rate and articulation rate of normal elderly people," Graduate program in speech and language pathology, Master these, Yonsei University, 2011.

[7] W. J. Ryan, J.William, "Acoustic aspects of the aging voice", Journal of Gerontology, vol.27, no.2, pp.265-268, 1972. doi : 10.1093/geronj/27.2.265

[8] Y.H. Kim. "Geriatric speech. plenary session IV," Yonsei University College of Medicine, Otolaryngology clinic. pp.205-207, 2003.

[9] W.H. Manning, K.L.Monte, "Fluency breaks in older speakers: implications for a model of stuttering throughout the life cycle," Journal of fluency disorders. Vol.6, no.1, pp.35–48, 1981. doi : 10.1016/0094-730x(81)90029-2

[10] J.D. Harnsberger, R. Shrivastav, R. Brown, W.S. Rothman, H. Hollien, "Speaking rate and fundamental frequency as speech cues to perceived age," Journal of voice, vol.22, no.1, pp.58-69, 2008. doi : 10.1016/j.jvoice.2006.07.004

[11] H.Y. Pyo, H.S. Shim, "Paralytic disorder words (dysarthria) for improving the clarity of research trends: A Literature Review," Special Education, vol.4, no.1, pp.35-50, 2005

[12] M. Richardson, M. Hwang, A, Acero, X.Huang, "Improvements on speech recognition for fast talkers," Eurospeech, pp.411-414, 1999.

[13] S. Kwon, S. Kim, J. Choeh. "Preprocessing for elderly speech recognition of smart devices," Computer Speech & Language. vol.36, pp.110-121, 2016. doi : 10.1016/j.csl.2015.09.002

[14] A. Aniruddha, M. Mathew, S. Amantula, C. Sekhar, "Gammatone wavelet Cepstral Coefficients for robust speech recognition," TENCON 2013, pp.1-4, 2013. doi : 10.1109/TENCON.2013.6718948

[15] W. Verhelst, M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," Acoustics, Speech, and Signal Processing(ICASSP), vol.2, pp. 554-557, 1993. doi: 10.1109/icassp.1993.319366

[16] W. Verhelst, "Overlap-add methods for time-scaling of speech. Speech Communication," vol.30, no.4, pp.207-221, 2000. doi: 10.1016/s0167-6393(99)00051-5

[17] C. d'Alessandro, "Time-frequency speech transformation based on an elementary waveform representation. Speech communication," pp.419-431, 1990. doi: 10.1016/0167-6393(90)90018-5

[18] D. Henja, B. Musicus "The solafs time-scale modification algorithm," Technical Report of BBN, 1991.

[19] S. Kwon, 'Voice-driven sound effect manipulation" International Journal of Human-Computer Interaction, pp.373–382, 2012. doi : 10.1080/10447318.2011.595359

[20] S. Dusan, L.R. Rabiner, "On the relation between maximum spectral transition positions and phone boundaries," INTERSPEECH, pp.17-21, 2006. doi : 10.1109/TSP.2006.885780