A Method of Evaluating Korean Articulation Quality for Rehabilitation of Articulation Disorder in Children

Keonsoo Lee[1] and Yunyoung Nam[2*]
[1]Convergence Institute of Medical Information Communication Technology and Management, Soonchunhyang University, Asan, Republic of Korea
[2]Department of Computer Science and Engineering, Soonchunhyang University, Asan, Republic of Korea
[e-mail: lks7256@hanmail.net, ynam@sch.ac.kr]
*Corresponding author: Yunyoung Nam

Received December 26, 2019; revised May 19, 2020; accepted June 18, 2020; published August 31, 2020

Abstract

Articulation disorders are characterized by an inability to achieve clear pronunciation due to misuse of the articulators. In this paper, a method of detecting such disorders by comparing to the standard pronunciations is proposed. This method defines the standard pronunciations from the speeches of normal children by clustering them with three features which are the Linear Predictive Cepstral Coefficient (LPCC), the Mel-Frequency Cepstral Coefficient (MFCC), and the Relative Spectral Analysis Perceptual Linear Prediction (RASTA-PLP). By calculating the distance between the centroid of the standard pronunciation and the inputted pronunciation, disordered speech whose features locates outside the cluster is detected.

89 children (58 of normal children and 31 of children with disorders) were recruited. 35 U-TAP test words were selected and each word's standard pronunciation is made from normal children and compared to each pronunciation of children with disorders. In the experiments, the pronunciations with disorders were successfully distinguished from the standard pronunciations.

Keywords: Articulation Disorder, LPCC, MFCC, RASTA-PLP, U-TAP

This research was supported by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (P0012724, The Competency Development Program for Industry Specialist) and the Soonchunhyang University Research Fund.

http://doi.org/10.3837/tiis.2020.08.006 **ISSN : 1976-7277**

1. Introduction

Speech articulation disorders are mainly resulted from the damage of sound production [1]. Typical symptoms include slurred and/or indistinct speech, and substituted sounds. Rehabilitation involves speech therapy such as physical exercises for the speech muscles, pronunciation practice, and speech drills to promote clear expression [2, 3]. However, rehabilitation is critically dependent on the quality of the speech pathologist [4]. The relationship between therapists and patients is important, and therapists must be able to distinguish between clear and slurred, indistinct, vague, and/or nebulous pronunciation. To distinguish “definitely good” from “clearly bad” pronunciation than to distinguish “somewhat poor” from “normal” pronunciation. There is an ambiguity between slightly bad and good which are relatively differently defined according to the pathologists [5]. Therapist experience and talent are critical for speech rehabilitation in children.

In this paper, a method, by which pronunciation quality in Korean children with articulation disorders can be evaluated and treated based on quantification of articulation, is proposed. The basic idea of the proposed method is that there is a standard pronunciation for each word. And pronunciations with disorders are different from the standard pronunciation. So, this method defines the standard pronunciation of words and determines the disordered pronunciation by calculating the distance from the standard pronunciation. As this proposed method is confined to the Korean language, it is not guaranteed to be used in other languages. Every language exhibits unique pronunciation rules. Therefore, the criteria for “good pronunciation" in Korean are not transferable to other languages. There are four basic features of the Korean language [6]. First, the Korean language has 24 atomic letters (14 consonants and 10 vowels) and 16 complex letters (5 consonants and 11 vowels). Second, each syllable has three parts, of which the first ("onset") is any of 19 consonants, the second ("nucleus") is any of 21 vowels, and the third ("coda") is any of 27 consonants. The coda is not always required; possible coda candidates are identified by reference to the atomic consonants, some of which cannot serve as onsets. In the coda, different consonants are pronounced similarly. Words that sound the same are written differently and have different meanings. Intonation is not a factor in modern Korean, unlike ancient Korean and some contemporary dialects. In general, regional accents should be ignored when evaluating possible articulation disorders.

The proposed method is valid only for children aged 4 to 12 years. Pronunciation of children is different from that of adults. Therefore, the standard pronunciation made from the children cannot be used as a standard pronunciation of adults for evaluating articulation quality [7].

This paper is organized as follows. Section 2 introduces the background of the proposed method. Section 3 describes the proposed methods. Section 4 shows the experimental results, and Section 5 provides a discussion and conclusions.

2. Background

2.1 The structure of voice signals

Voice signals have three main components, the first of which is noise. As sounds are transferred through the air, the original signals become contaminated by the noise that can be reduced using various methods [8]. If the noise is defined, it may be readily removed.

However, if the noise is not defined, identification and filtering thereof are essential; the adaptive Weiner filter (AWF) [9], spectral subtraction [10], and gamma-tone filter (GTF) [11] can be employed to this end. The AWF detects noise by reducing the difference between the filtered and intended signals. Spectral subtraction assumes that non-speech signals are noise; thus, they are removed. The GTF uses a filter bank to remove signals unrecognized by the human cochlea.

The second component is speech information, i.e., the words were actually spoken. Regardless of the speaker, this is identical to a given utterance. The third component is speaker information. For any given sentence, pronunciation may differ among speakers; pronunciation is individual and can therefore can be used to identify a subject [12]. When evaluating articulation quality, speech information must be extracted from voice signals.

2.2 Features for Speeches

Various features can be extracted from speech signals, of which the Linear Predictive Cepstral Coefficient (LPCC), Mel-frequency Cepstral Coefficient (MFCC), and Relative Spectral Analysis Perceptual Linear Prediction (RASTA-PLP) are the most widely used. An LPCC is a linear predictive coefficient (LPC) in the cepstral domain, acquired by subjecting the logarithm of the speech power spectrum to inverse discrete Fourier transformation (IDFT). As the LPC is acquired via speech autocorrelation, the LPCC is calculated by subjecting the logarithm of the smoothed auto-regressive speech power spectrum to IDFT [13]. An MFCC is similar to an LPCC, but the former uses a Mel-frequency" scale ranging from 300 to 3,400 Hz [14]. This encompasses the range of human hearing; removal of high-frequency signals reduces noise. The RASTA-PLP enhances a PLP by applying a RASTA filter [15]. A PLP is a form of LPC involving perceptual processing, critical-band spectral resolution, an equal-loudness curve, and the intensity-loudness power law. After RASTA filtering, slow-channel speech variations are suppressed, rendering subsequent PLP more robust. In our method, LPCC, MFCC, and RASTA-PLP are all applied.

2.3 Similarity calculations

Similarity reflects between-object distance in feature space; the nearer the object, the greater the similarity. Euclidean distance, a subset of Minkowski distance, often serves as the distance between objects [16]. If m is 1, the Manhattan distance is used, while if m is 2, the Euclidean distance is employed.

$$\textbf{Minkowski Distance} = \sqrt[m]{\sum_{i=1}^{n} (|a_i - b_i|)^m} \tag{1}$$

In this study, we used the Euclidean distance to determine the extent to which a given pronunciation differs from standard pronunciation.

2.4 Speech recognition

Speech recognition systems translate spoken language into text [17]. Speech recognition systems have three components, the first of which is a model generation for the target language. Such models subsume acoustic and language sub-models. The second component extracts features from speech. The third component classifies patterns ("speech matches"). As speech proceeds, features are extracted from the signals, and compared to those of the acoustic model a set of phonemes is then retrieved. The language model is then used to yield the most reliable text. The Gaussian mixture model (GMM) and the hidden Markov model (HMM) are conventionally employed to decode features extracted from signals, thus yielding text [18]. The GMM defines the feature pattern, while the HMM indicates the statistical relationships among patterns. GMMs are currently being replaced by deep neural network (DNN)-based pattern-matching methods [19]. Recently, DNNs have also replaced the HMM, creating an end-to-end acoustic model [20]. Various speech recognition systems are available; commercial products include the Samsung Bixby [21], Google Assistant [22], Amazon Alexa [23], Apple Siri [24], and Microsoft Cortana [25]. The most popular non-commercial products include the Mozilla Common Voice Project [26], Kaldi [27], CMUSphinx [28], HTK [29], and Julius [30]. This study is not concerned with speech recognition or translation; rather, we aimed to determine pronunciation quality. As the text to which the speech is matched is known, speech recognition is unnecessary. However, our method can be used to recognize the speech of children with articulation disorders.

3. Proposed Method

3.1 Evaluation of pronunciation quality

It is more difficult to define good articulation than to identify improved articulation based on a set of pronunciations. Intuitively, pronunciation that is easy to recognize is optimal [31]. To quantify speech quality, a standard pronunciation for each word is first identified; the quality of a subject's pronunciation is given by its distance from the standard pronunciation.

3.2 Feature Extraction

Raw signals are preprocessed in three steps. In the first step, background noise is removed using a bandpass filter. The typical frequency of a child’s voice is 250–300 Hz [32]; a filter that passes signals of 200–350 Hz is thus used. In the second step, silence is removed and speech amplitude increased. In the third step, pronunciation length is adjusted by resampling as shown in **Fig. 1**.

After preprocessing, the MFCC, LPCC, and RASTA-PLP features are extracted from the signal. **Fig. 2** shows the features extracted from the signals as shown in **Fig. 1**.

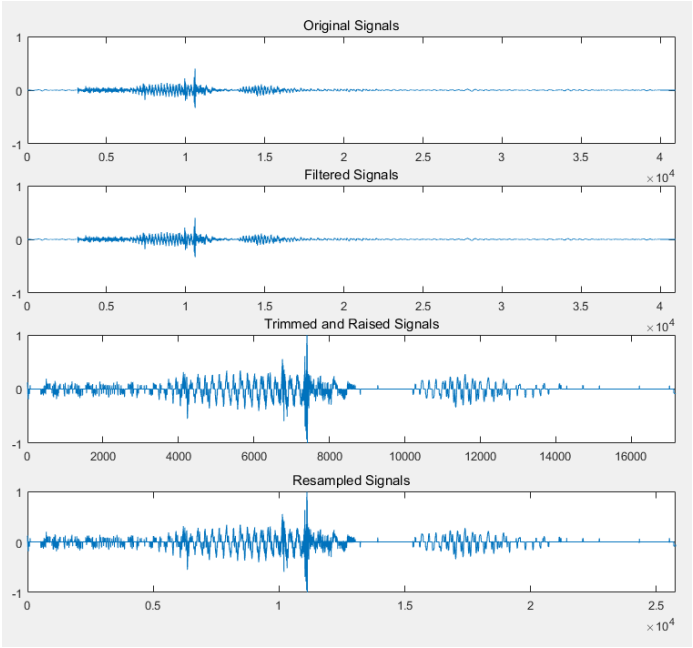

Fig. 1. Preprocessing steps of speech signals

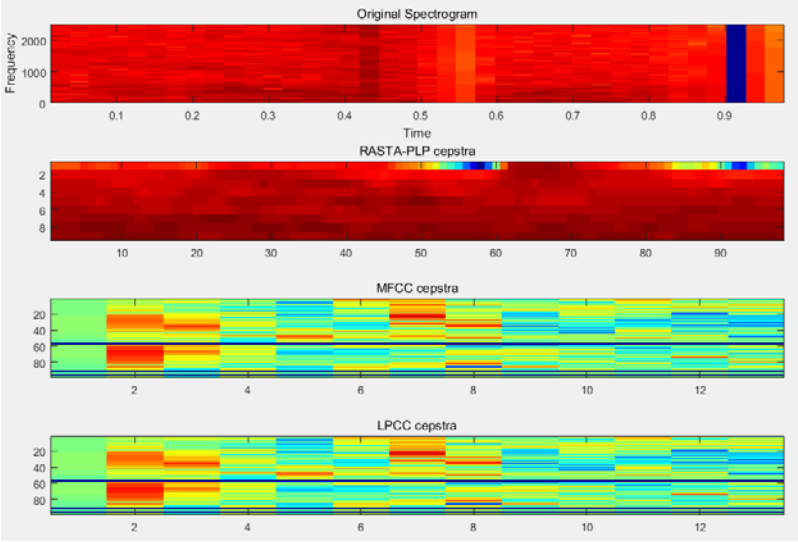

Fig. 2. Three features extracted from the speech signals

3.3 Quality evaluation

Pronunciation quality is given by the distance between the actual and standard pronunciation; the latter is the centroid of the pronunciation cluster of normal children. Disordered pronunciations are scattered around the normal cluster and the distance from the centroid denotes the pronunciation quality. The X-means clustering method identifies the number of normal pronunciation clusters. If multiple clusters are evident, the highest-density

cluster is taken to reflect normal pronunciation; the other clusters are semi-normal pronunciations. As shown in **Fig. 3**, the area of each cluster is divided into three regions, of which A is the area within which the distance to the centroid is < 50% that to the boundary. Region B is the area within which the distance to the centroid is > 50% but < 80% of the distance to the boundary. Region C is the remaining area.

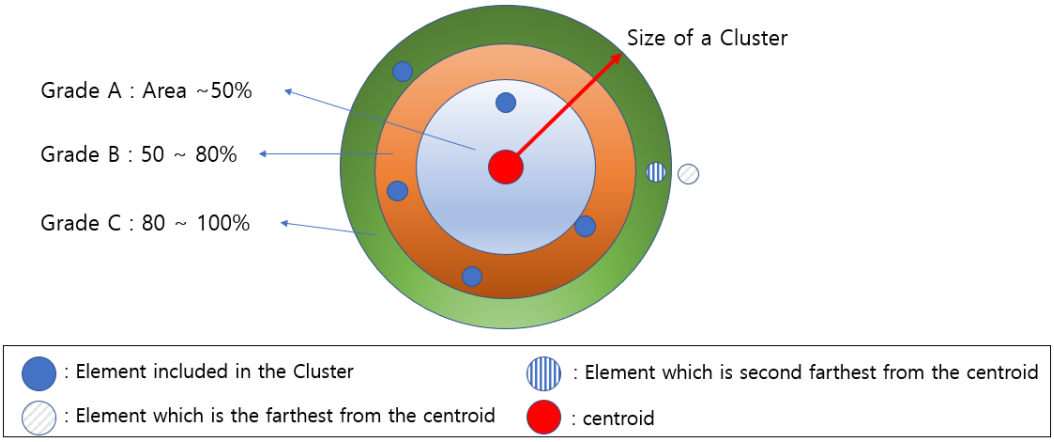

Fig. 3. Features of defining a cluster

The distance from the centroid to the boundary (the "size") of a cluster is calculated using Equation (2). Cluster density is calculated by dividing the number of elements by the cluster distance using Equation (3). A cluster containing more elements within a short distance is optimal.

$$\mathbf{SoC} = \mathbf{Euclidean}\left(\mathbf{Centroid}, \mathbf{E_{fp}}\right) + \frac{\mathbf{1}}{\mathbf{2}}\left(\mathbf{E_{ff}}, \mathbf{E_{sf}}\right) \tag{2}$$

$$\mathbf{Density} = \frac{\boldsymbol{NoC}}{\boldsymbol{SoC}} \tag{3}$$

SoC : Size of the given cluster
E_{ff} : Element which is the farthest from the centroid
E_{sf} : Element which is the second farthest from the centroid
NoC : The number of elements in a given cluster

A, B, and C serve as ordinal pronunciation quality grades; pronunciations lying closer to the centroid are of higher quality. Pronunciation quality is nominally graded as "normal" and "semi-normal"; the semantics of relationships among multiple clusters are not explored in this study.

4. Experiments

4.1 Data Collection

89 children aged 3–14 years who visited the Department of Otorhinolaryngology, Soonchunhyang University Hospital, Cheonan, Korea are recruited for evaluating the proposed method. Of these 89 children, 58 did not have an articulation disorder, so their

speeches were used to establish standard pronunciations. The 31 children with articulation disorders were aged 3–6 years; their speeches were used to validate the standard pronunciations. The Urimal Test of Articulation and Phonology (U-TAP) [33], which is widely used in Korea, was employed to acquire the pronunciations. The U-TAP evaluates both word and sentence pronunciations. 35 words with 43 phonemes from the U-TAP were selected for this experiment. **Table 1** shows the meanings and pronunciations.

Table 1. Words used to collect pronunciations

No	*Word, Pronunciation, Meaning*	*No*	*Word, Pronunciation, Meaning*	*No*	*Word, Pronunciation, Meaning*
1	나무 , /na-mu/, Tree	13	냉장고, /n æ ŋ-ʒa ŋ-go/, Refrigerator	25	식빵, /sik-fa ŋ/, Bread
2	목도리, /mok-do-ri/, Muffler	14	단추, /dan- tʃu/, Button	26	입술, /ip-sul/, Lips
3	꽃, /ggot/, Flower	15	공, /go ŋ/, Ball	27	포크, /fɔrk/, Fork
4	김밥, /gim-bab/, Rice roll	16	가방, /ga-ba ŋ/, Bag	28	짜장면, /tza- ʒa ŋ-mi ʌ n/, Jjajangmyeon
5	바지, /ba-ʒi/, Trousers	17	똥, / ðo ŋ/, Poop	29	싸움, / θa-um/, Fight
6	사탕, /sa-ta ŋ/, Candy	18	쌀, / θal/, Rice	30	로봇, /rou-bət/, Robot
7	풍선, /pu ŋ-s ən/, Balloon	19	책상, /tʃæk-sa ŋ/, Desk	31	팝콘, /pap-cɔrn/, Popcorn
8	국자, /guk-ʒa/, Scoop	20	자동차, /ʒa-doŋ-tʃa/, Car	32	접시, /ʒʌb-si/, Plate
9	토끼, /to-kki/, Rabbit	21	해바라기, /h æ-ba-ra-gi/, Sunflower	33	기차, /gi-tʃa/, Train
10	코끼리, /ko-kki-ri/, Elephant	22	연필, /i ən-pil/, Pencil	34	그네, /gɯ-n ɛ/, Swing
11	호랑이, /ho-ra ŋ-i/, Tiger	23	빨간색, /pal-gan-s æk/, Red	35	짹짹, / tz æk- tz æk/, Chirping
12	라면, /ra- mən/, Noodle	24	참새, / tʃam-s æ/, Sparrow		

The sentence pronunciation tests, which are provided in U-TAP, were not used in the experiment because analyzing the pronunciation of sentences has different criteria and variations compared to the words. A method for exploring Korean pronunciation based on the rules of words interactions, and analyzing sentences will be executed in future work.

4.2 Experimental plan

The proposed method is composed of two sub-processes as shown in **Fig. 4**.

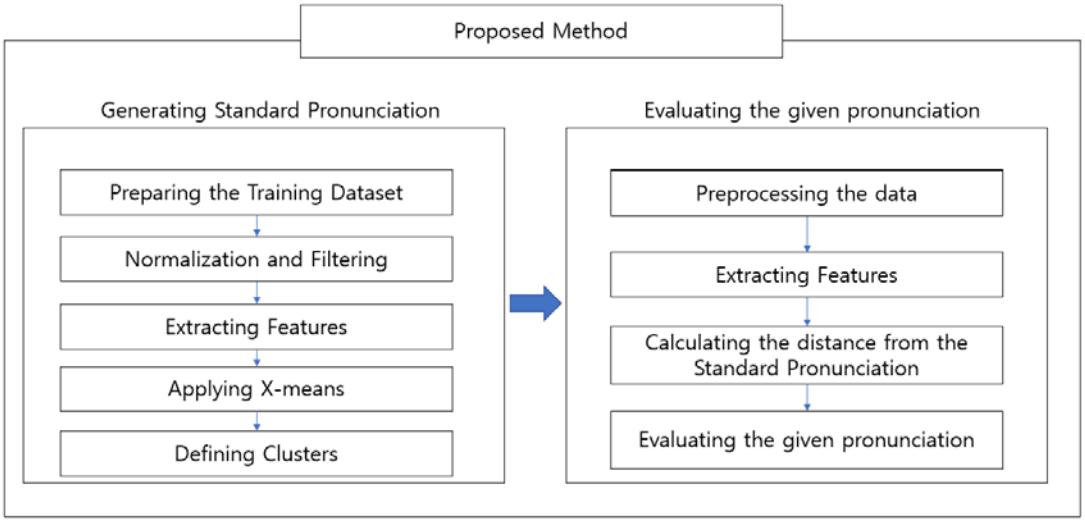

Fig. 4. The experimental process

Standard pronunciations were defined based on the speech of 58 normal children. First, the articulations for 35 U-TAP words were extracted and grouped. This was followed by signal preprocessing and filtering; word amplitude, length, and sampling rate were then normalized. In terms of amplitude normalization, the minimum and maximum strength of the original signals were set from -1 to 1, respectively. The average length of each pronunciation was then determined; shorter or longer pronunciations were re-sampled to ensure that each pronunciation had the same number of samples. Low-quality pronunciations were removed. The quality of the normalized data was determined according to the agreement between the Google Speech-to-Text (STT) engine and a normal volunteer (a male college student). As subtle pronunciation classification is not required, a non-specialist with good hearing can achieve accurate classification after minimal training. In most cases, the STT engine and human evaluator were in the agreement. All human-recognized "good" pronunciations were accepted by the STT engine. However, some "poor" STT pronunciations were accepted by the human evaluator. Only agreed-upon pronunciations were used; 10~20% of records were thus discarded. The discard rate increased as the word pronunciation difficulty increased. The MFCC, LPCC, and RASTA-PLP features were then extracted. The pronunciations were clustered using the X-means algorithm. Finally, clusters defining the standard pronunciations of each word were created.

Using the defined standard pronunciations, test pronunciations were evaluated. First, each pronunciation was preprocessed. Then, the word amplitude, length, and sampling rate were normalized. The three features such as MFCC, LPCC, and RASTA-PLP were extracted and the distances from the centroid clusters of the standard pronunciations were calculated. If the distance was less than the cluster size, pronunciation was considered normal and graded as A, B, or C. If the distance was greater than the cluster size, the articulation disorder was diagnosed.

4.3 Results

Table 2 shows the densities of the standard pronunciation clusters and the numbers of clusters for the 35 words listed in **Table 1**. Cluster density was calculated by dividing the cluster volume by the number of elements, which is described in Equation (3). Each volume was calculated assuming that the cluster was spherical and had a radius as determined by equation (2). The MFCC, LPCC, and RASTA-PLP were equally weighted when calculating the Euclidean distance, and were normalized to minimize bias.

Table 2. The numbers of normal pronunciation clusters for each word and the densities of the principal clusters

No	*# of clusters*	*Density*	*No*	*# of clusters*	*Density*	*No*	*# of clusters*	*Density*
1	1	6.7406	13	1	8.5463	25	2	16.0254
2	2	8.7389	14	2	7.5874	26	1	4.0932
3	1	9.1535	15	1	8.4444	27	1	13.5801
4	1	8.7988	16	1	13.7794	28	2	12.1338
5	1	5.3482	17	1	9.8298	29	1	7.0005
6	2	8.7615	18	1	5.2272	30	3	4.0125
7	1	10.6526	19	1	9.3619	31	2	12.8818
8	1	5.9304	20	1	15.671	32	1	6.4487
9	2	6.6831	21	3	3.9095	33	1	9.7175
10	2	12.0989	22	1	12.7335	34	1	7.2004
11	2	5.8148	23	2	6.2169	35	2	15.6369
12	1	5.3702	24	1	7.0723			

After clustering, the speech of 31 children with articular disorders was evaluated in terms of distance from the standard pronunciation centroids as shown in **Table 3**. As all tested pronunciations were disordered, all pronunciations should be classified accordingly, i.e., as "true negatives". Articulations classified as normal corresponded to "false positives". No true-positive or false-negative articulations were found.

For most words, articulation disorders were correctly classified. However, polysyllabic words without codas were sometimes misclassified. Syllables without codas were easier to pronounce, while words with more syllables became increasingly difficult to pronounce. Pronunciation difficulty varied according to the consonant type.

Table 3. The classification of articulation disorders

No	*True Negative*	*False Positive*	*No*	*True Negative*	*False Positive*	*No*	*True Negative*	*False Positive*
1	21	10	13	31	0	25	31	0
2	31	0	14	31	0	26	31	0
3	31	0	15	28	3	27	31	0

4	31	0	16	31	0	28	31	0
5	28	3	17	31	0	29	31	0
6	31	0	18	31	0	30	31	0
7	31	0	19	31	0	31	31	0
8	31	0	20	31	0	32	31	0
9	31	0	21	31	0	33	31	0
10	31	0	22	31	0	34	24	7
11	31	0	23	31	0	35	31	0
12	27	4	24	31	0			

4.4 Discussion

In this study, standard pronunciations for 35 U-TAP words were defined from the pronunciations of 58 normal children. Comparing the standard pronunciations, articulation disorders were classified. The pronunciations of 31 children, who were diagnosed by the pathologists, were accurately detected.

The speeches of children aged 3–14 years were used to generate standard pronunciations. However, speech pathologists have found that articulation maturity increases with age [34]. Variation in vocal organ maturity may explain the existence of multiple clusters. This feature explains why some words have multiple clusters and other words have single cluster. Not only the difficulty in pronunciation resulted from the number of syllables but also the variation of pronunciation according to the maturity of the vocal organ makes wider margin for classifying the disordered pronunciation. Therefore, it is recommended to divide children into age- and gender-specific subgroups, and future studies should include more subjects. Nevertheless, the proposed method can reliably classify disordered speech. If the pronunciation of older children affected the standard pronunciation of younger children, the pronunciations of the younger children with articulation disorders were not included in the cluster. Thus, although pronunciation quality may vary by age, disordered speech can nonetheless be detected.

5. Conclusion

In this paper, pronunciation quality is determined by comparing standard and individual pronunciations. To define pronunciations, 875 records were collected from 89 children. 58 children do not have an articulation disorder; thus, their speech was used to define standard pronunciations. The speeches of 31 children were employed to validate these pronunciations. The MFCC, LPCC, and RASTA-PLP features were extracted from each record. The X-means algorithm was used to define clusters of normal pronunciation. By calculating the distances between cluster centroids and individual pronunciations, the disordered pronunciation, which locates outside of the standard pronunciation cluster, is detected. Standard pronunciations were derived for 35 U-TAP words. Relatively short, simple words in the U-TAP word list formed single clusters. Meanwhile, longer and more difficult words formed several clusters.

From the experiment, the pronunciations of children with articulation disorders were correctly classified using the proposed method. However, standard pronunciations with

multiple clusters, which means that the word is difficult to renunciate and has variations according to the maturity of the vocal organ, seems to have too wide margin for detecting disordered pronunciation. In order to reduce the margin and make more accurate cluster, normal children were further subdivided using more detailed fractionation in ages and gender. In future work, standard pronunciations by age and gender will be defined. Then, the proposed method could be used for evaluating articulation and communication disorders with words as well as sentences.

References

[1] P. Grunwell, "The phonological analysis of articulation disorders," *British Journal of Disorders of Communication*, vol. 10, no. 1, pp. 31-42. 1975. Article (CrossRef Link).

[2] B. Dodd, Z. Hua, S. Crosbie, S., A. Holm, A., and A. Ozanne, *Diagnostic Evaluation of Articulation and Phonology—US Edition (DEAP)*, Pearson, San Antonio, TX, 2009.

[3] B. Dodd, and A. Bradford, "A comparison of three therapy methods for children with different types of developmental phonological disorder," *International Journal of Language & Communication Disorders*, vol. 35, no. 2, pp. 189–209, 2000. Article (CrossRef Link).

[4] D. S. Borys, and K. S. Pope, "Dual relationships between therapist and client: A national study of psychologists, psychiatrists, and social workers," *Professional Psychology: Research and Practice*, vol. 20, no. 5, pp. 283-293, 1989. Article (CrossRef Link).

[5] V. Hodge, and J. Austin, "A survey of outlier detection methodologies," *Artificial intelligence review*, vol. 22, no. 2, pp. 85-126. 2004. Article (CrossRef Link).

[6] J. J. Song, *The Korean language: Structure, use and context*, Routledge, 2006. Article (CrossRef Link).

[7] L. L. Olson, and S. Jay Samuels, "The relationship between age and accuracy of foreign language pronunciation," *The Journal of Educational Research*, vol. 66, no. 6, pp. 263-268, 1973. Article (CrossRef Link).

[8] A. Ravishankar, S. Anusha, H. K. Akshatha, A. Raj, S. Jahnavi, and J. Madhura, "A survey on noise reduction techniques in medical images," in *Proc. of 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, vol. 1, pp. 385–389, April 20-22, 2017. Article (CrossRef Link).

[9] A. Yelwande, S. Kansal, and A. Dixit, "Adaptive wiener filter for speech enhancement," in *Proc. of 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC)*, pp. 1–4, August 17-19, 2017. Article (CrossRef Link).

[10] S. V. Vaseghi, "Spectral Subtraction," *Advanced Signal Processing and Digital Noise Reduction*, pp. 242–260, 1996. Article (CrossRef Link).

[11] K. Odugu and B. M. S. S. Rao, "New speech enhancement using Gamma tone filters and Perceptual Wiener filtering based on sub banding," in *Proc. of 2013 INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING AND COMMUNICATION (ICSC)*, pp. 236–241, December 12-14, 2013. Article (CrossRef Link).

[12] S. Kanrar, "Speaker Identification by GMM based i Vector," *arXiv:1704.03939 [cs]*, April 2017.

[13] J. D. Markel and A. H. Gray, *Linear prediction of speech*, Springer- Verlag, New York, 1976. Article (CrossRef Link).

[14] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients in the power spectrum," in *Proc. of 2001 IEEE international conference on Acoustics, Speech, and Signal Processing*, 2001. Article (CrossRef Link).

[15] J. Koehler, N. Morgan, H. Hermansky, H. G. Hirsch, and G. Tong, "Integrating RASTA-PLP into speech recognition," in *Proc. of ICASSP `94. IEEE international conference on Acoustics, Speech, and Signal Processing*, 1994. Article (CrossRef Link).

[16] Z. Li, Q. Ding, and W. Zhang, "A Comparative Study of Different Distances for Similarity Estimation," *Intelligent Computing and Information Science*, pp. 483–488, 2011. Article (CrossRef Link).

[17] J. A. Bilmes, “Graphical Models and Automatic Speech Recognition,” *Mathematical Foundations of Speech and Language Processing*, pp. 191–245, 2004. Article (CrossRef Link).
[18] S Renals, N Morgan, H Bourlard, M Cohen, and H Franco, "Connectionist probability estimators in HMM speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 161-174, 1994. Article (CrossRef Link).
[19] G. Hinton, L. Deng. D. Yu, G Dahl, A. R. Mohamed, N, Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury, and T. Sainath “Deep Neural Networks for Acoustic Modeling in Speech Recognition,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, 2012. Article (CrossRef Link).
[20] A. Graves, A. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proc. of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645–6649, May 26-31, 2013. Article (CrossRef Link).
[21] “Samsung Bixby: Your Personal Voice Assistant | Samsung US,” Samsung Electronics America. [Online]. Available at: /us/explore/bixby/.
[22] “Google Assistant,” Google Assistant. [Online]. Available: https://assistant.google.com /.
[23] “Ways to Build with Amazon Alexa,” [Online]. Available: https://developer.amazon.com/alexa
[24] “Siri,” Apple. [Online]. Available: https://www.apple.com/siri/
[25] “Personal Digital Assistant - Cortana Home Assistant - Microsoft,” Microsoft Cortana, your intelligent assistant. [Online]. Available: https://www.microsoft.com/en-us/cortana.
[26] “Common Voice by Mozilla,” [Online]. Available: https://mzl.la/voice. [Accessed: 05-Feb-2019].
[27] “Kaldi ASR,” [Online]. Available: http://kaldi-asr.org/.
[28] N. Shmyrev, “CMUSphinx Open Source Speech Recognition,” CMUSphinx Open Source Speech Recognition. [Online]. Available: http://cmusphinx.github.io/.
[29] “HTK Speech Recognition Toolkit,” [Online]. Available: http://htk.eng.cam.ac.uk/.
[30] “Open-Source Large Vocabulary CSR Engine Julius,” [Online]. Available: http://julius.osdn.jp/en_index.php.
[31] J. D. O'Connor, *Better English Pronunciation*, Cambridge University Press, 1980.
[32] E. J. Hunter, “A comparison of a child’s fundamental frequencies in structured elicited vocalizations versus unstructured natural vocalizations: A case study,” *International Journal of Pediatric Otorhinolaryngology*, vol. 73, no. 4, pp. 561–571, 2009. Article (CrossRef Link).
[33] Y. Kim, H. Park, J. Kang, J. Kim, M. Shin, S. Kim, J Had, “Validity and Reliability Analyses for the Development of Urimal Test of Articulation and Phonology-2,” *Commun Sci Disord*, vol 23, no. 4, pp. 959–970, 2018. Article (CrossRef Link).
[34] E. M. Griebeler, N. Klein, and P. M. Sander, “Aging, Maturation and Growth of Sauropodomorph Dinosaurs as Deduced from Growth Curves Using Long Bone Histological Data: An Assessment of Methodological Constraints and Solutions,” *PLoS One*, vol 8, no. 6, pp.1-17, 2013. Article (CrossRef Link).

Keonsoo Lee He received the M.S. and Ph.D. degrees in computer engineering from Ajou University, Korea, in 2004 and 2013, respectively. He is currently a Research Professor at the Medical Information Communication Technology, Soonchunhyang University, Asan, Korea. His research area includes artificial intelligence, knowledge representation, and multi-agent system.

Yunyoung Nam received the B.S., M.S., and Ph.D. degrees in computer engineering from Ajou University, Korea in 2001, 2003, and 2007 respectively. He was a Senior Researcher with the Center of Excellence in Ubiquitous System, Stony Brook University, Stony Brook, NY, USA, from 2007 to 2010, where he was a Postdoctoral Researcher, from 2009 to 2013. He was a Research Professor with Ajou University, from 2010 to 2011. He was a Postdoctoral Fellow with the Worcester Polytechnic Institute, Worcester, MA, USA, from 2013 to 2014. He was the Director of the ICT Convergence Rehabilitation Engineering Research Center, Soonchunhyang University, from 2017 to 2020. He has been the Director of the ICT Convergence Research Center, Soonchunhyang University, since 2020, where he is currently an Assistant Professor with the Department of Computer Science and Engineering. His research interests include multimedia database, ubiquitous computing, image processing, pattern recognition, context-awareness, conflict resolution, wearable computing, intelligent video surveillance, cloud computing, biomedical signal processing, rehabilitation, and healthcare systems.