PAPER

Speech Clarity Index (Ψ): A Distance-Based Speech Quality Indicator and Recognition Rate Prediction for Dysarthric Speakers with Cerebral Palsy

Prakasith KAYASITH[†], ***Student Member and*** **Thanaruk THEERAMUNKONG**[†a)], ***Member***

SUMMARY It is a tedious and subjective task to measure severity of a dysarthria by manually evaluating his/her speech using available standard assessment methods based on human perception. This paper presents an automated approach to assess speech quality of a dysarthric speaker with cerebral palsy. With the consideration of two complementary factors, *speech consistency* and *speech distinction*, a speech quality indicator called *speech clarity index* (Ψ) is proposed as a measure of the speaker's ability to produce consistent speech signal for a certain word and distinguished speech signal for different words. As an application, it can be used to assess speech quality and forecast speech recognition rate of speech made by an individual dysarthric speaker before actual exhaustive implementation of an automatic speech recognition system for the speaker. The effectiveness of Ψ as a speech recognition rate predictor is evaluated by *rank-order inconsistency*, *correlation coefficient*, and *root-mean-square of difference*. The evaluations had been done by comparing its predicted recognition rates with ones predicted by the standard methods called the *articulatory* and *intelligibility tests* based on the two recognition systems (HMM and ANN). The results show that Ψ is a promising indicator for predicting recognition rate of dysarthric speech. All experiments had been done on speech corpus composed of speech data from eight normal speakers and eight dysarthric speakers.

key words: *speech disorder, dysarthric speech recognition, speech assessment, speech quality index, recognition rate prediction*

1. Introduction

Dysarthria is a term representing a group of speech disorder in which the transmission of messages controlled by the motor movements for speech is interrupted. To provide suitable means of rehabilitation or assistance for people in this group, we need a reliable method to identify their problems and evaluate their severity. To this end, contemporarily there are two common tests based on perceptual analysis, namely an *articulatory test* and an *intelligibility test*. The articulatory test has been widely used as a clinical tool to evaluate the severity of a speaker and to identify the errors of dysarthric speech. The test is quite rigid in the sense that the speech is evaluated in the terms of linguistic aspect. Naturally, the result relies on perceptions of clinicians (speech therapists or pathologists) and their knowledge and experience about the assessment method. Therefore it is very subjective and occasionally inconsistent among different judges, especially when it is performed by clinicians who may have different common knowledge and training [1].

As the other standard test, the intelligibility evaluation is normally performed by a group of non-hearing impaired listeners (usually consists of 6 - 12 persons) instead of speech specialists or trained listeners. The main objective of the test is to measure a level of understanding between a speaker and a listener. Therefore the absolute correctness (or clearness) of speech in terms of linguistics is not important as long as the message (or information) from the speaker can be understood by the listener. The results of the test come from the average value of all assessments done by each listener. There are some attempts to measure intelligibility in some other means rather than a simple human speech-perception. For example, Power and Braida [2] at MIT focused on consistency measurement to predict intelligibility degraded by noise. Another work had been done by Bodt *et al.* [3] in order to express intelligibility as linear combination of the four main speech factors; voice quality, articulatory, nasality, and prosody. However, each factor still was judged by human.

While both articulatory and intelligibility tests are widely used, they are time-consuming and subjective to human perception [3], [4]. Several cases pointed out some disagreements in the results obtained from different evaluators [5]. Moreover, the results may not explicitly represent the level of severity in terms of speech signal processing such as consistency in speech. In several cases, severe dysarthric speech may be completely unintelligible to unfamiliar listeners; however, to a familiar communication partner, the same speech seems to be intelligible. A possible interpretation of this phenomenon is that, people learn to understand speech even it does not follow general linguistic rules of that language. It can be observed that when people continue making conversation with a dysarthric speaker, they try to grasp common acoustic (signal) characteristics of a same word/phrase and at the same time attempt to find acoustic (signal) characteristic difference among different word/phrases in order to adaptively learn to recognize dysarthric speech. Based on this observation, an assumption has been made; given one dysarthric speaker, if his/her familiar communication partner can recognize (or learn to recognize) his/her speech, some modern speech processing techniques (i.e., speech recognition) should be able to learn

Manuscript received June 11, 2008.
Manuscript revised October 1, 2008.

[†] The authors are with Information and Computer Technology School, Sirindhorn International Institute of Technology, Thammasat University, Thailand.

a) E-mail: thanaruk@siit.tu.ac.th

DOI: 10.1587/transinf.E92.D.460

to recognize those patterns as well. In the past, several studies [6]–[8] showed that with carefully designed; a dysarthric speaker can be benefited from incorporating a speech recognition system into assistive devices.

Motivated by these works, a potential assumption is that even most dysarthric speakers tend to produce speeches that are not linguistically correct but if they can produce consistent speech for a same word and generate distinctive speeches for different words, high quality of speech can be expected, resulting in a high recognition rate when a recognition system is made for them.

To this end, this paper proposes a speech quality indicator called speech clarity index (Ψ), which can be used to indicate quality of speech done by a speaker and applied to predict outcome performance of speech recognition systems before system implementation. As the investigation domain, we focuses on dysarthric speakers with cerebral palsy, ones with damage in their cerebral palsy. In the rest, a conceptual idea and mathematical terms are presented in Sect. 2. The experimental details including characteristics of the subjects, the speech corpus, and the evaluation methods are shown in Sect. 3. Section 4 describes the experimental results and some discussions on them. Section 5 gives a conclusion and future works.

2. Speech Clarity Index (Ψ)

So far there have been several existing standard methods of speech assessment developed for different purposes, such as clinical test, speech rehabilitation, and assessment of speech comprehensive quality, but they are usually subjective and time-consuming tasks. There have been still few works on automatic speech assessment [9], [10]. Carmichael *et al.* [9] presumed a correlation between probabilistic likelihood scores from HMM-based speech recognition system and the results of intelligibility test from eight listeners in order to formulate the conversion algorithm for intelligibility prediction. Lingyun *et al.* [10] applied dynamic time wrapping (DTW) and the Itakura-Saito (IS) distortion measure to evaluate the distortion of speech quality by comparing it to healthy speech done by normal speakers. However, both works aimed to measure the difference between disordered speech and normal speech.

This paper presents an automated speech assessment method that measures the quality of the individual's speech without comparing to the others'. For this purpose, a distance-based speech quality indicator called a *speech clarity index* (Ψ) is proposed.

Basically, Ψ is designed to indicate speech clarity (quality) of an individual using two complementary concepts of *speech consistency* (SC) and *speech distinction* (SD). The former depicts the similarity of speech signals of the same word/phrase produced several times by the speaker. The latter indicates the dissimilarity of speech signals of different words/phrases produced by the speaker. A graphical representation of SC and SD is shown in Fig. 1. Conceptually, SC means the closeness of speech signals depicted by the average distance within a group ($D11$, $D22$, and $D33$); whereas SD represents the distinction of the signals from different groups expressed by the average distance that separates each group from the others ($D12$, $D23$, and $D13$). Then Ψ is designed as the combination of those two speech factors. A high value of Ψ is achieved when there is high speech consistency within a group and high speech distinction among different groups.

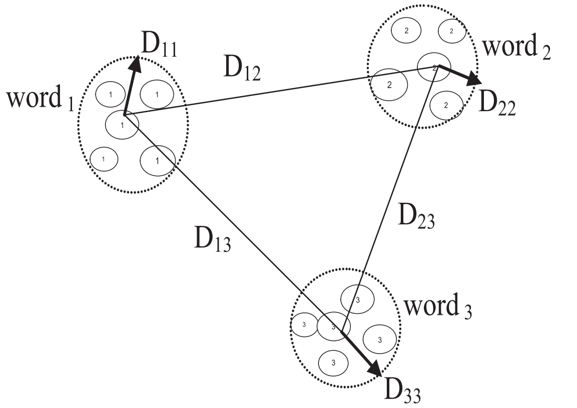

Fig. 1 Graphical expression of speech consistency and speech distinction.

To define SC and SD mathematically, the following two problems need to be clarified. The first problem involves how to encode a speech signal into a form of feature vector sequences. This process is also known as feature extraction. The second problem concerns how to compare two sequences of feature vectors, so as to derive a distance between any two signals.

As for feature extraction, each of two signals will be first separated into a sequence of smaller frames (typically 25 ms width with 10 ms interval, i.e., 15 ms overlapping). Then, each frame will be encoded into a speech feature vector by a standard feature extraction method. In general, the definition of features is arbitrary. It can be defined basing on the characteristics of interest such as intensity of energy and its variations, the frequency spectrums, or the formants analyses during that time frame. In this work, we choose *12-th order Mel Frequency Cepstral Coefficients* ($MFCC$) [11] with their first and second derivatives, since they are commonly used in general.

As for feature vector comparison, to solve a problem of time variation of speech samples, a so-called *dynamic time warping* (DTW) can be applied [12], [13]. Proposed in this work, a high variation of dysarthric speech signals is controlled under three constraints, that is (i) *adaptive endpoint selection*, (ii) *adaptive slope constraint*, and (iii) *accumulated penalty coefficient for repeated frames*. The first constraint stands for determining the synchronized endpoints of the two signals. Two possible endpoint zones have been set; one at the beginning and the other at the end (shading areas in Fig. 2). Then within each endpoint zone, the best match pair of frames (minimum distance) is selected to be the new starting point or ending frames. For example in Fig. 2, the

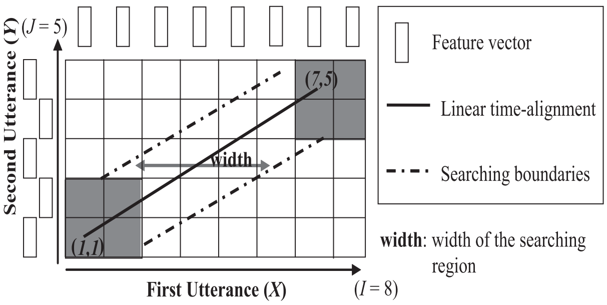

Fig. 2 Comparison of two sequences of feature vectors (X and Y) using DTW.

original starting and ending frames for X and Y are [1,1] and [8,5], respectively. By the first constraint, the newly selected endpoints are set to [1,1] and [7,5], as shown in the figure.

The second constraint states that the optimal path should be located along the linear time-alignment path (a solid line in Fig. 2); and the searching boundaries and the slope should be varied according to duration difference between the two utterances. The width of the searching region (width) lied in the principle direction (i.e., the direction of the longer utterance) is proportional to the difference of I and J (i.e., $width = |I - J|$). Then, the wrapping function for searching bounds above and beneath the linear time-alignment path (the dash lines in Fig. 2) was generated with the window size of $|I - J|$ and a slope of J/I.

The last constraint is to determine a suitable *penalty coefficient* to cope with a high variation of speech signals. In our experiments, the penalty coefficient is varied from 0 to 5 in order to search for the optimal final penalty score which comes from the multiplication of the number of repeated frames and the penalty coefficient. By preliminary experiments, the most suitable coefficient is 2.5, and this value is used for all experiments in this paper. As for *distance-based* feature vector comparison, the difference between two feature vectors (frames) $\vec{x} = [x_1, x_2, \ldots, x_p]$ and $\vec{y} = [y_1, y_2, \ldots, y_p]$ is defined as Euclidean distance, $d(\vec{x_i}, \vec{y_j})$.

As a formal description, two speech signals that we are going to compare can be transformed to two sequences of feature vectors, $X = \{\vec{x_i} | i \in \{1, 2, \ldots, I\}\}$ and $Y = \{\vec{y_j} | j \in \{1, 2, \ldots, J\}\}$. Note that X and Y consist of I and J frames, respectively. Here, let L be a set of all possible alignments. Each alignment $l = \{(i_k, j_k) | k \in \{1, \ldots, K\}, 1 \leqslant i_k \leqslant I, 1 \leqslant j_k \leqslant J, (i_k = i_{k-1} \text{ or } i_{k-1} + 1), (j_k = j_{k-1} \text{ or } j_{k-1} + 1), i_K = I, j_K = J\}$, expresses a mapping between two sequences X and Y. Based on a certain alignment, the DTW distance between these two utterances ($DTW[X, Y]$) is defined in Eq. (1).

$$DTW[X, Y] = \min_{l \in L} \left(\left(\sum_{(i,j) \in l} d(\vec{x_i}, \vec{y_j}) \right) + p(l) \right) \tag{1}$$

Here, $p(l)$ is a *penalty score* defined by the penalty coefficient multiplied by the number of repeated frames in the alignment l. A set of repeating frames is a subset of the alignment such that the pair is not diagonal matching, denoted by $\{(i_r, j_r) | r \in 1, 2, \ldots, K, (i_r, j_r) \in l, (i_r = i_{r-1} \text{ or } j_r = j_{r-1})\}$. Using the above basic calculation, three measures, SC, SD, and Ψ, are formulated as follow.

As the first measure, speech consistency (SC) of a speaker is derived from how different the pronouncing utterances of the same word are. Assume that there are N representative words (later indexed by $w \in \{1, 2, \ldots, N\}$) covering the major basic phonetic pronunciations in one's language. For each word, M utterances pronounced by the same speaker are collected. To figure out SC, the consistency of speech is calculated by measuring distances of those pronouncing words. To do this, we first measure consistency of all pronunciations of each word and then combine the results of all words to gain overall evaluation.

In the first step, for each word indexed by w with M utterance samples, the average within-word distance d^w can be calculated by Eq. (2).

$$d^w = \frac{1}{{}^{M}C_2} \sum_{i=1}^{M} \sum_{j=i+1}^{M} DTW[X_i^w, X_j^w] \tag{2}$$

where X_i^w and X_j^w are the sequences of feature vectors, which encode the i-th and the j-th utterances of the word w, consecutively. Here, there are ${}^{M}C_2 (= \frac{M(M-1)}{2})$ pairs of feature vectors comparison. As for a word w, the distance d^w indicates deviation among M pronunciations of the word. A small distance means high consistency of several pronunciations of the same word produced by a speaker.

In the second step, the average distance (d_{avg}) of the N representative words is derived from Eq. (3).

$$d_{avg} = \frac{1}{N} \sum_{w=1}^{N} d^w \tag{3}$$

Finally, speech consistency can be defined as the inverse of d_{avg} as follow.

$$SC = \frac{1}{d_{avg}} \tag{4}$$

As the second measure, speech distinction (SD) of a speaker is used to represent the average distance between pronunciations of different words uttered by a speaker. To measure this, first a suitable representative utterance of each word is chosen. For the N representative words, we will obtain N representative utterances. Second, the between-word distance of two words is defined by the distance between the representative utterances of these two words. With ${}^{N}C_2 (= \frac{N(N-1)}{2})$ pairs of words, SD is derived by calculating the average distance among the between-word distances of these pairs. The following describes the detail of calculation. In the step of selecting the representative utterance for the word w denoted by T_w. As a simple approach, our method considers the utterance that has the minimum sum of distances away from the other utterances of the same word, mathematically defined in Eq. (5).

$$T^w = \underset{T \in X^w}{\operatorname{argmin}} \left(\sum_{j=1}^{M} DTW[T, X_j^w] \right) \tag{5}$$

where $X^w = \{X_1^w, X_2^w, \ldots, X_M^w\}$ is a set of collected utterances of the w-th word. Then SD is calculated according to Eq. (6).

$$SD = \frac{1}{{}^{N}C_2} \sum_{i=1}^{N} \sum_{j=i+1}^{N} DTW[T^i, T^j] \tag{6}$$

While SC indicates how the speaker pronounces a same word consistently, SD displays how he/she pronounces different words distinctly. A speaker with high speech consistency and high speech distinction is implied to have high speech clarity. To fulfill a meaningful speech evaluation, both parameters have to be taken into account. Integrating SC and SD, the speech clarity index (Ψ) is formulated to express the ability of a speaker in producing speech clearly, and then defined by $SC \times SD$. The index can be used to indicate the quality of speech made by a speaker and then to predict recognition performance gained from an automatic speech recognition system.

3. The Experiments

3.1 The Word Set and Its Speech Corpus

To avoid transition effect that may occur in a multi-syllable word, we select a monosyllable target word for each most commonly used Thai phoneme, to construct a speech corpus for our experiment. As our selection criterion, every target word must be inferred easily from a simple picture in order to collect natural speech and eliminate the problem of literacy in children or CP subject. Under this constraint, among the 70 Thai phonemes, only 67 phonemes are chosen to construct the so-called phoneme-test set of 67 target words for conducting Thai phoneme error analysis. Here, two vowels [u : a, ua] and one consonant cluster [tr] are excluded since they are rarely used and hard found in any simple word.

In corpus construction, due to the physical limitation of each CP-dysarthric child, only five speech samples can be recorded for each target words. The speech samples are recorded under semi-controlled environment, i.e. in a quiet room with the door closed but no additional sound proof materials. A dynamic headset microphone (Shure Model SM2) is used at a position of approximately 1.5 cm. from the right side of the speaker's mouth. During a recording process, speech stimuli (the target pictures) are presented to a speaker on a computer screen. The speaker is instructed to utter each word separately using habitual pronunciation. If the speaker cannot utter the target word, the system will give an example pronunciation of the word. Moreover, the picture will be repeated after a pre-setup order. The speech samples are recorded with a 16-bit A/D converter at a sampling rate of 16 kHz.

Table 1 Demographic and characteristic of the Cerebral palsy children.

Code	Age	Sex	Cerebral Palsy	Dysarthria
DF01	11	F	Athetoid	Hypokinetic
DF02	12	F	Flaccid Hemiplegia	Flaccid
DF03	7	F	Spastic Diplegia	Spastic
DF04	12	F	Athetoid	Hypokinetic
DM01	12	M	Spastic Diplegia	Spastic
DM02	13	M	Athetoid	Hypokinetic
DM03	10	M	Athetoid	Hypokinetic
DM04	14	M	Athetoid	Hypokinetic

Table 2 Speech severity level for *A*-score and *I*-score.

Severity Level	Severity Score
Very Severe	0.00 – 0.40
Severe	0.41 – 0.60
Moderate	0.61 – 0.80
Mild	0.81 – 0.95
Normal	0.96 – 1.00

3.2 The Subjects

The subjects are of two groups. The normal group is composed of eight normal speakers; four adults (23–36 years old) and four children (7–12 years old). The dysarthric group consists of eight cerebral palsied dysarthric (CP-dysarthric) children (7–14 years old). Both groups are formed with the balance number of males and females. All CP-dysarthric children are students from the Srisungwan compulsive school, a school for children with disabilities. All subjects have been tested for hearing acuity within normal range and no mental retardation problem (IQ more than 70 or above). As for a subject code, each subject is referred by a 4-digit code. The first digit represents the subject group; where 'D', 'A', and 'N' stand for 'dysarthric speaker', 'normal adult speaker' and 'normal children speaker', respectively. The second digit expresses the sex of the subject; where 'M' means 'male' and 'F' is for 'female'. The last two digits are a running number of the subject. For example, DM02 means the second CP-dysarthric male speaker.

Table 1 displays the demographic and characteristic of the CP children evaluated by experts. As references, all subjects have been assessed by two standard tests, an articulatory test and an intelligibility test. In the articulatory test, an *articulatory score* (later denoted by *A*-score), assigned to each speaker judged by two experts, is the ratio of the number of correctly pronounced phonemes to the total number of pronounced phonemes. The criterion is a modification of *"Percentage of Correct Consonants (PCC)"* [14], [15]. In the intelligibility test, a speaker's speech had been assessed by 12 non-hearing impaired listeners under three sessions; *word transcription*, *multiple choices*, and *rating scale*. The scores gained from all listeners' evaluations for these three sessions are summed up and averaged to display the *intelligibility score* (later denoted by *I*-score) for the speaker. Given the criterion for determining the level of speech severity as shown in Table 2, the result of mapping of *A*-score and

Table 3 *A*-score, *I*-score and the severity level of each subject.

	Articulatory Test		Intelligibility Test	
Code	***A*-score**	**Severity**	***I*-score**	**Severity**
AF01	1.00	Normal	0.99	Normal
AF02	1.00	Normal	0.99	Normal
AM01	1.00	Normal	1.00	Normal
AM02	1.00	Normal	0.99	Normal
NF01	0.97	Normal	0.97	Normal
NF02	0.91	Mild	0.94	Mild
NM01	0.97	Normal	0.91	Mild
NM02	0.91	Mild	0.94	Mild
DF01	0.63	Moderate	0.53	Severe
DF02	0.49	Severe	0.39	Very Severe
DF03	0.66	Moderate	0.77	Moderate
DF04	0.51	Severe	0.48	Severe
DM01	0.56	Severe	0.41	Severe
DM02	0.69	Moderate	0.63	Moderate
DM03	0.69	Moderate	0.77	Moderate
DM04	0.63	Moderate	0.68	Moderate

I-score onto a speech severity level for each subject is shown in Table 3.

3.3 The Two Speech Recognition Systems as References

Two speaker-dependent speech recognition systems, Hidden Markov Model (HMM) and Artificial Neural Network (ANN), are used as the references. As for the HMM toolkit, the HTK version 3.2.1 is employed. This toolkit is developed at the Machine Intelligence Laboratory (formerly known as the Speech Vision and Robotics Group) of the Cambridge University Engineering Department [16]. In our experiments, each phone is characterized by 5-state HMM with a Gaussian state distribution. A pronunciation dictionary is employed to map word pronunciations to a sequence of phones for each word. Since the task in this work focuses on isolated words, a high-level language model, e.g. bigram/trigram, is not used

The NICO (Neural Inference COmputation) toolkit developed by the department for Speech, Music and Hearing at KTH, Stockholm [17] is used for our ANN system. A three-layer (i.e. input, hidden, and output) feed-forward network is chosen for the experiments. While the number of input units depends on the number of input features, that of output units and the hidden nodes are set to 67 (the number of words in the corpus) and 100, respectively. For the input, we apply linear selection of 15 frames per word. Therefore, the number of input nodes is 540 (15 frames×36 features). All features are normalized to the range of -1.0 to 1.0. The network was trained by the standard back-propagation method with a random initial weight between -1.0 and 1.0.

As stated in Sect. 3.1, five speech samples have been collected for each word per subject. Each simple is encoded with the 12^{th} order Mel Frequency Cepstral Coefficient ($MFCC$) with their first and second derivatives. To evaluate recognition rates obtained from HMM and ANN, we select the 5-fold cross validation (in this case, equivalent to leave-one-out). Four samples for each word are used for training and the rest one is used for testing. This process repeats five times and the averaged recognition rate is calculated.

3.4 Evaluation Methods

Compared to *A*-score and *I*-score, the usability of Ψ as the index to predict speech recognition rate for a dysarthric speaker is evaluated with the results from the recognition systems, HMM and ANN. In this work, we evaluate Ψ using three different measures. As the first measure, *rank-order inconsistency* (*ROI*) is calculated as follow. Given the set of subjects, we directly compare the results of Ψ (also *A*-score and *I*-score) to the recognition rate from the recognition system (either HMM or ANN), in terms of performance order. As the second measure, the *Pearson's correlation coefficient* (R^2) between Ψ (also *A*-score and *I*-score) and the recognition rate (of HMM or ANN) is calculated to indicate how much Ψ correlates with the recognition result. For the last measure, *root-mean-square of difference* (Δ_{rms}) is computed as follow. First, a regression is performed to find the relation between Ψ (also *A*-score and *I*-score) and the recognition performance (of HMM or ANN). As the result, a function for predicting the recognition performance is generated. These functions are used to predict recognition performance from Ψ (also *A*-score and *I*-score). Using the predicted recognition rates and the actual recognition rates (of HMM or ANN), Δ_{rms} is calculated. Since different recognition methods may give different results, for each measure we calculate a margin (an acceptable bound) by considering the difference between the performances of HMM and ANN. The margin is used to determine whether the difference between the predicted recognition rates from Ψ (also *A*-score and *I*-score) and the recognition rate is acceptable or not. The details of these three measures are given in the next subsections.

3.4.1 Rank-Order Inconsistency (*ROI*)

Given two different methods that rank a set of N objects and may result with different orders, the mismatch between their rankings can be counted using the techniques called pairwise comparison [18]. This technique states that the two methods have no rank mismatch on an object pair, say o_i and o_j, if both methods agree that o_i is better than o_j or vice versa. Otherwise, there is a rank mismatch. To this end, the number of rank mismatches between the two methods, say X and Y, called a mismatch score, is defined as $M(X, Y)$ by Eq. (7).

$$M(X, Y) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} |\delta(x_i, x_j) - \delta(y_i, y_j)| \tag{7}$$

where x_i and x_j are the respective rank of the *i*-th and the *j*-th objects based on the method X while y_i and y_j are the respective rank of the *i*-th and the *j*-th objects based on the method Y. $\delta(a, b)$ is the *mismatch function*. It returns 1 when a less than b otherwise 0.

The Rank-Order Inconsistency (*ROI*) shown in Eq. (8) is calculated by dividing the mismatch score in Eq. (7) with the mismatch score of the worst case, i.e. all data in one method is arranged in the reverse order compared to the other method, i.e. $N(N-1)/2$. Note that *ROI* ranges between 0 (the identical order) to 1 (the reverse order).

$$ROI(X,Y) = \frac{2 \times M(X,Y)}{N(N-1)} \tag{8}$$

In our experiment, among N speakers to be evaluated, x_i and x_j are speech-quality ranks of the i-th and the j-th speakers suggested by a speech quality index (Ψ, *A*-score or *I*-score), respectively. While y_i and y_j represent recognition ranks of the i-th and the j-th speakers when their speeches are recognized by a recognition system, HMM or ANN. Based on this, *ROI* is calculated to indicate rank mismatch between the speech quality index (Ψ, *A*-score or *I*-score) and speech recognition rate (HMM or ANN).

3.4.2 Pearson's Correlation Coefficient (R^2)

As another measure, one can investigate correlation of the results from two different methods (X and Y) using the Pearson's correlation coefficient (R). Under the assumption of Gaussian distribution in both methods, the correlation coefficient can be written as

$$R = \frac{\sum_{i=1}^{N}(x_i - x_{avg})(y_i - y_{avg})}{(N-1)\sigma_x\sigma_y} \tag{9}$$

where N is a number of objects in comparison, x_i and y_i represent the respective results of the i^{th} objects given by the methods X and Y, x_{avg} and y_{avg} are the respective average given by the methods X and Y, and σ_x and σ_y are the respective standard deviations given by the methods X and Y. In our experiment, N is a number of speakers whose speech we are going to evaluate, x_i, x_{avg} and σ_x are the speech quality of each speaker, the average speech quality of all speakers, and the standard deviation of speech quality of all speakers, and y_i, y_{avg} and σ_y are the recognition rate of each speaker, the average recognition rate of all speakers, and the standard deviation of recognition rates of all speakers. The speech quality can be measured by means of either of Ψ, *A*-score or *I*-score while the recognition rate can be determined by either HMM or ANN.

3.4.3 Root-Mean-Square of Difference (Δ_{rms})

Besides the rank mismatch and the correlation analysis, it is possible to compare the results from two different methods (X and Y) using the root-mean-square of difference (Δ_{rms}). The formula is quite straightforward as shown in the following equation.

$$\Delta_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \tag{10}$$

where N is a number of objects in comparison, x_i and y_i represent the respective results of the i-th objects given by the methods X and Y.

Unlike the rank mismatch and the correlation analysis, the root-mean-square of difference is quite sensitive to the size difference between the results from the two methods. This implies that the results from the two methods need a form of normalization to the same scale before comparison. For this purpose, we use a linear regression process to map the value of a speech quality index (Ψ, *A*-score or *I*-score) to a value of recognition rate (HMM or ANN) before calculating Δ_{rms}. For this work, N is a number of speakers whose speech we are going to evaluate, x_i is a recognition rate predicted from a speech quality index (Ψ, *A*-score or *I*-score) for each speaker, and y_i is the recognition rate of each speaker given by either HMM or ANN.

4. Results and Discussions

Table 4 shows the results of HMM recognition rate (SRR_{HMM}), ANN recognition rate (SRR_{ANN}), articulatory test (*A*-score), intelligibility test (*I*-score), SC, SD and Ψ in the cases of normal speakers and dysarthric speakers. The mean value of Ψ for the normal group is 1.5 ($\sigma = 0.11$) while it is 1 ($\sigma = 0.10$) for the CP-dysarthric group. For a specific speaker, a large value of SC means he/she has high consistency in his/her pronunciations of the same word while a large value of SD shows he/she can make a clear distinction (conspicuously separate) in pronouncing different words. Unsurprisingly, a normal speaker gains a high speech recognition rate (SRR) while a dysarthric speaker tends to obtain a lower speech recognition rate. The tendency is also the same for the proposed speech clarity index and the two standard assessment methods.

Figure 3 shows a graphical relation between SC^{-1} and SD, and a distance between these two measures reflects a value of Ψ. For a normal group, SD of each speaker is always at the outer loop (higher value) and keeps a considerable distance away from SC^{-1} (with an approximate ratio of 1.5). As for CP-dysarthric group, in most cases, SC^{-1} and SD values are close to each other (with an approximate ratio of 1.0). In some particular cases such as DF01, DF04, and DM02, this ratio (Ψ) are less than 1. These figures imply the characteristic of highly-overlapped speech signals produced by those dysarthric subjects. As the consequence, performance of a speech recognition system tends to be low for these cases. On the contrary, a speaker with a Ψ value higher than 1, has a potential to gain a high recognition rate.

As mentioned in the last paragraph of Sect. 2, Ψ represents the ability that a speaker can produce consistent speech when he/she pronounces a word several times, and produce distinct speech when he/she utters different words. When a speaker has a high value of Ψ, it implies that the speaker can control his speaking ability to produce a clear speech. In this case, a high recognition rate can be expected. Naturally, we can expect a high correlation between Ψ and recognition rates.

Table 4 Experimental results of normal group (AF01-NM02) and CP-dysarthric group (DF01-DM04).

Code	SRR_{HMM}	SRR_{ANN}	*A*-score	*I*-score	SC	SD	Ψ
AF01	0.99	0.93	1.00	0.99	0.0172	88.82	1.52
AF02	0.99	0.97	1.00	0.98	0.0185	84.48	1.57
AM01	0.98	0.95	1.00	1.00	0.0178	93.66	1.67
AM02	0.98	0.97	1.00	0.99	0.0162	88.76	1.44
NF01	0.98	0.95	0.97	0.97	0.0195	79.94	1.56
NF02	0.92	0.84	0.91	0.94	0.0179	82.58	1.48
NM01	0.95	0.84	0.97	0.91	0.0143	89.49	1.28
NM02	0.94	0.91	0.91	0.94	0.0184	84.53	1.56
DF01	0.38	0.38	0.63	0.53	0.0078	110.51	0.86
DF02	0.49	0.54	0.49	0.39	0.0115	87.31	1.00
DF03	0.77	0.65	0.66	0.77	0.0079	132.21	1.04
DF04	0.51	0.47	0.51	0.48	0.0096	98.47	0.94
DM01	0.55	0.53	0.56	0.41	0.0090	112.10	1.01
DM02	0.49	0.37	0.69	0.63	0.0058	154.48	0.89
DM03	0.72	0.60	0.69	0.77	0.0112	93.84	1.05
DM04	0.75	0.77	0.63	0.68	0.0127	91.34	1.16

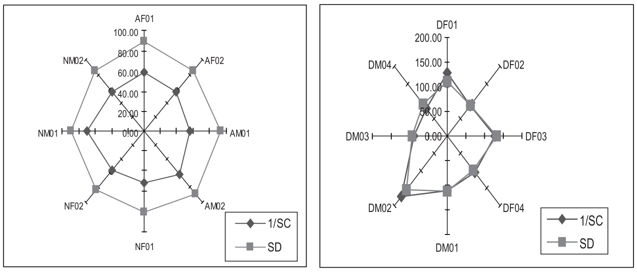

Fig. 3 SC^{-1} and SD diagram of a normal group (left) and a CP-dysarthric group (right).

Table 5 Evaluation results of Ψ (five and three samples/word), *A*-score, and *I*-score on three measurements (*ROI*, R^2, and Δ_{rms}), when SRRs obtained from HMM and ANN are used as the references. The numbers in the parentheses stand for the cases that only dysarthric speeches are considered.

Reference	Method	*ROI*	R^2	Δ_{rms}
HMM	**Ψ**	0.1000	0.8616	0.0800
	(5 samples)	(0.1071)	(0.7383)	(0.0692)
	Ψ	0.1500	0.8318	0.0882
	(3 samples)	(0.1071)	(0.6573)	(0.0792)
	A-**score**	0.1000	0.8283	0.0891
		(0.3214)	(0.1652)	(0.1237)
	I-**score**	0.0917	0.8953	0.0696
		(0.2857)	(0.5503)	(0.0908)
ANN	Ψ	0.0917	0.9132	0.0629
	(5 samples)	(0.1071)	(0.9618)	(0.0247)
	Ψ	0.1083	0.8827	0.0731
	(3 samples)	(0.1071)	(0.8970)	(0.0406)
	A-**score**	0.1500	0.7674	0.1030
		(0.5000)	(0.0056)	(0.1262)
	I-**score**	0.1583	0.7812	0.0999
		(0.4643)	(0.1943)	(0.1136)

*The numbers in each parenthesis is the result when 'only dysarthric speakers' are considered

4.1 Evaluations Using *ROI*, R^2, and Δ_{rms}

Using speech recognition rates of HMM and ANN as references, a series of experiments and evaluations have been explored. Based on speeches of all speakers (dysarthric and normal speakers), the evaluation on the proposed index (Ψ) and the other two standard indicators (*A*-score and *I*-score) have been done using three different measures. Their results are shown in Table 5. The numbers in the parentheses stand for the cases that only dysarthric speeches are considered. The performance of Ψ on full samples (five speech samples per word) is indicated. To investigate the robustness of Ψ, an additional exploration is made to find the performance of Ψ when three samples per word are considered. Their results are shown in the first two rows for HMM and ANN.

We can observe that Ψ shows the lowest rank-order inconsistency (*ROI*) with the ANN recognition system, while the intelligibility test gains the lowest ROI with HMM. When the calculation of Ψ is based on a smaller training dataset (three samples per word), ROI increases slightly from 0.10 and 0.09 to 0.15 and 0.10 for HMM and ANN, respectively. This indicates that Ψ is still a good indicator when a small dataset is applied. In cases of only dysarthria's speech, Ψ seems to be the best indicator that matches with recognition rates of both HMM and ANN while the A-score and I-score obtained high ROIs, implying that they are not good measures for evaluating dysarthria's speech.

For the evaluation using correlation, Ψ achieves the highest correlation (R^2) of 0.91 with ANN and obtains a correlation of 0.86 with HMM. For the articulator test, the R^2 between *A*-score and the recognition rates of HMM and ANN, are 0.82 and 0.76, respectively. For the intelligibility test, *I*-score seems to gain the highest correlation of 0.89 with the HMM, while it gains a correlation of 0.78 with the ANN. For Ψ calculated from three samples per word, it also gained a same figure of results, showing it is still a good indicator for ANN but is relatively good for HMM.

The result implies that Ψ matches better to the performance obtained from the ANN model than to that gained from the HMM model. One possible reason is that HMM is a time-dynamic pattern recognition model that also includes an additional language model into the system which may not be reflected in the property of Ψ. On the other hand, ANN and Ψ are similar at the points of time alignment and pattern comparison, i.e., no additional language model. As one more interesting observation, the *I*-score seems to gain the highest correlation with the HMM result. One potential explanation is that the intelligibility test is based on human perception, where a listener may implicitly exploit some kinds of language models during their judgment. This characteristic is similar to what HMM uses for recognition.

An additional experiment on only the CP-dysarthric speaker group, we found out that there is a very high correlation (R^2 = 0.96) between Ψ and the ANN results. The correlation between Ψ and the HMM results is also high (R^2 = 0.74). This correlation coefficient is higher than the case of considering all speakers (normal and CP-dysarthria). One possible explanation for this case is that a relation of Ψ and the recognition rate for the normal speaker group is more complex. It may not be linear as we observe in the CP-dysarthric group. The results of normal speakers are quite sensitive to the change of Ψ. The results suggest a non-linear relation at the high Ψ region. If all speakers (normal and dysarthria) are involved, the best equation should be a kind of logarithmic relation, instead of a linear relation. However, at the low Ψ region (speakers with dysarthria), the linear relation can be assumed. Therefore, a recognition rate prediction function of each index can be derived from the linear regression method.

As for the evaluation on prediction error using Root-Mean-Square of Difference, Ψ gains the lowest prediction error of 6.29% for ANN and 8.00% for HMM, respectively. For *A*-score, they are 8.91% (HMM) and 10.30% (ANN). They are 6.96% (HMM) and 9.99% (ANN) for *I*-score. For the CP-dysarthric speaker group, we found out Ψ achieves the lowest prediction error for both HMM and ANN, compared to *A*-score and *I*-score. It is also a good indicator when a smaller data set (three samples per word) is used for calculation. Among the two standard assessment methods, *I*-score seems to be better than *A*-score in terms of predicting the recognition rate of both HMM and ANN.

4.2 Application of Ψ for Speech Severity Index for Dysarthria

According to the previous section, our proposed index seems to be good indicator for predicting recognition rates, compared to the standard assessment indicators. As another interesting application, it is possible to apply Ψ as an alternative speech severity index, in complimentary with the articulatory score and the intelligibility score. While the articulatory test is subjective to correctness of pronunciation and the intelligibility test focuses on message understanding between speakers and listeners, Ψ shows the performance of how well a speaker can produce consistent speech for a word and distinct speech for different words. Here, it is possible to evaluate speech disorder severity using this index. To evaluate the severity of dysarthric speech disorder with respect to speech recognition criteria, the linear regression function between Ψ and the averaged recognition rate (between two platforms, HMM and ANN) is calculated. From the regression, the result shows the value of correlation coefficient (R^2) at nearly 0.90. To set the guideline for severity index, we use the same severity criteria as shown in Table 2. The severity score calculated from the regression function for each speaker are shown in Table 6.

Table 6 Speech severity evaluation by Ψ, *A*-score, and *I*-score of each dysarthric speaker.

Code	*A*-score	*I*-score	Ψ
DF01	Moderate	Severe	Severe
DF02	Severe	Very Severe	Severe
DF03	Moderate	Moderate	Moderate
DF04	Severe	Severe	Severe
DM01	Severe	Severe	Severe
DM02	Moderate	Moderate	Severe*
DM03	Moderate	Moderate	Moderate
DM04	Moderate	Moderate	Moderate

*Different evaluation from *A*-score and *I*-score

The severity evaluations using an articulatory test and an intelligibility test are also given in the same table as references. The results from both articulatory test and intelligibility are consensus in most cases excepted for DF01 and DF02. As for the Ψ evaluation, there is only one case (DM02) that the suggested severity does not match with any suggested severity from these two standard tests. This result indicates that these three methods share some common decisions on dysarthric speech evaluation.

5. Conclusion and Future Works

Speech assessment for people with speech disorder is very important. The current methods of the articulatory test and the intelligibility test are subjective to individual expert and listener. In this work, it was shown that these standard methods that rely on human-perception do not effectively reflect the recognition rate result of modern speech recognition. Therefore, it may not suitable to use these standard tests to estimate the effectiveness of a speech recognition system on speech produced by a dysarthric speaker. In addition, both articulatory test and intelligibility test are a time-consuming and laborious task. This paper proposed the speech clarity index (Ψ) to predict the recognition rate gained from a speech recognition system. The predicted recognition rate can be served to decide whether the dysarthric speaker could be benefit from the speech recognition technology or not. It was shown that Ψ is a powerful indicator to predict speech recognition rate under three measures, rank-order inconsistency, correlation coefficient, and root-mean-square of difference. Compared to the HMM results, Ψ matches better with the ANN results since they share some common features of pattern comparison. On the other hand, the intelligibility test that seems related to the language model, may suit with the prediction of HMM recognition rates. By reducing the number of samples for calculating Ψ, this indicator is still good for predicting the recognition rate. As our future works, the research will focus on exploring more parameters such as a kind of overlapping factors between speech signals, the consistency of energy, time, and language models to improve the prediction of speech quality.

Acknowledgements

This research is supported by the Royal Golden Jubilee Ph.D. Program, Thailand Research Fund (TRF) under the contract number PHD/0267/2545.

References

[1] R.D. Kent, “Hearing and believing: Some limits to auditory-perceptual assessment of speech and voice disorders,” J. Speech and Hearing Disorders, vol.7, pp.7–23, 1996.

[2] M. Power and L. Braida, “Consistency among speech parameter vectors: Application to predicting speech intelligibility,” J. Acoust. Soc. Am., vol.100, no.6, pp.3882–3898, 1996.

[3] M. Bodt, M. Hernadez-Diaz, and P. Van De Heyning, “Intelligibility as a linear combination of dimensions in dysarthric speech,” J. Communication Disorders, vol.35, no.3, pp.283–292, 2002.

[4] J. Bernthal and N.W. Bankson, Articulation and phonological disorders, 3rd ed., Prentice Hall, Boston, 1993.

[5] R.D. Kent, G. Miolo, and S. Bloedel, “The intelligibility of children’s speech: A review of evaluation procedures,” American Journal of Speech-Language Pathology, vol.3, pp.81–95, 1994.

[6] J. Deller, D. Hsu, and L. Ferrier, “On the use of hidden Markov modeling for recognition of dysarthric speech,” Computer Methods and Programs in Biomedicine, vol.35, pp.125–139, 1991.

[7] A. Kotler and N. Thomas-Stonel, “Effects of speech training on the accuracy of speech recognition for an individual with speech impairment,” J. Augmentative and Alternative Communication, vol.12, pp.71–80, 1997.

[8] K. Rosen and S. Yampolsky, “Automatic speech recognition and a review of its functioning with dysarthric speech,” J. Augmentative and Alternative Communication, vol.16, pp.46–60, 2000.

[9] J. Carmichael and P. Green, “Revisiting dysarthria assessment intelligibility metrics,” 8th International Conference on Spoken Language Processing (ICSLP), pp.742–745, Jejunce, South Korea, 2004.

[10] L. Gu, J.G. Harris, R. Shrivastav, and C. Sapienza, “Disordered speech assessment using automatic methods based on quantitative measures,” EURASIP Journal on Applied Signal Processing, vol.2005, no.9, pp.1400–1409, 2005.

[11] P. Mermelstein, “Distance measures for speech recognition, psychological and instrumental,” in Pattern Recognition and Artificial Intelligence, ed. R. Chen, pp.374–388, Academic Press, 1976.

[12] F. Itakura, “Minimum prediction residual principle applied to speech recognition,” IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-23, no.1, pp.67–72, 1975.

[13] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-26, no.1, pp.43–49, 1978.

[14] L. Shriberg and J. Kwiatkowski, “Phonological disorders iii: A procedure for assessing severity of involvement,” J. Speech and Hearing Disorders, vol.47, no.3, pp.256–270, 1982.

[15] L. Shriberg, D. Austin, B.A. Lewis, J.L. McSweeny, and D. Wilson, “The percentage of consonants correct (pcc) metric. extensions and reliability data,” J. Speech, Language, and Hearing Research, vol.40, pp.708–722, 1997.

[16] HTK, HTK home (online), http://htk.eng.cam.ac.uk, 2008.

[17] NICO, NICO home (online), http://nico.nikkostrom.com, 2008.

[18] H. David, The method of paired comparisons, Oxford University Press, New York, 1988.

Prakasith Kayasith received a B.Sci (Physics) and a M.Sci (Computer) from Chulalongkorn University, Thailand. Thereafter, he received his M.Eng in Electrical Engineering from Old Dominion University, USA. He is currently a Ph.D. candidate in Information Technology at Sirindhorn International Institute of Technology, Thaialnd. His researches focus on clinical application of speech technology for person with disabilities.

Thanaruk Theeramunkong received a bachelor degree in Electric and Electronics, and master and doctoral degrees in Computer Science from Tokyo Institute of Technology in 1990, 1992 and 1995, respectively. His current research interests include data mining, machine learning, natural language processing, and information retrieval.