Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 540409, 12 pages
doi:10.1155/2009/540409

Research Article

Alternative Speech Communication System for Persons with Severe Speech Disorders

Sid-Ahmed Selouani,[1] Mohammed Sidi Yakoub,[2] and Douglas O'Shaughnessy (EURASIP Member)[2]

[1] *LARIHS Laboratory, Université de Moncton, Campus de Shippagan, NB, Canada E8S 1P6*
[2] *INRS-Énergie-Matériaux-Télécommunications, Place Bonaventure, Montréal, QC, Canada H5A 1K6*

Correspondence should be addressed to Sid-Ahmed Selouani, selouani@umcs.ca

Received 9 November 2008; Revised 28 February 2009; Accepted 14 April 2009

Recommended by Juan I. Godino-Llorente

Assistive speech-enabled systems are proposed to help both French and English speaking persons with various speech disorders. The proposed assistive systems use automatic speech recognition (ASR) and speech synthesis in order to enhance the quality of communication. These systems aim at improving the intelligibility of pathologic speech making it as natural as possible and close to the original voice of the speaker. The resynthesized utterances use new basic units, a new concatenating algorithm and a grafting technique to correct the poorly pronounced phonemes. The ASR responses are uttered by the new speech synthesis system in order to convey an intelligible message to listeners. Experiments involving four American speakers with severe dysarthria and two Acadian French speakers with sound substitution disorders (SSDs) are carried out to demonstrate the efficiency of the proposed methods. An improvement of the Perceptual Evaluation of the Speech Quality (PESQ) value of 5% and more than 20% is achieved by the speech synthesis systems that deal with SSD and dysarthria, respectively.

1. Introduction

The ability to communicate through speaking is an essential skill in our society. Several studies revealed that up to 60% of persons with speech impairments have experienced difficulties in communication abilities, which have severely disrupted their social life [1]. According to the Canadian Association of Speech Language Pathologists & Audiologists (CASLPA), one out of ten Canadians suffers from a speech or hearing disorder. These people face various emotional and psychological problems. Despite this negative impact on these people, on their families, and on the society, very few alternative communication systems have been developed to assist them [2]. Speech troubles are typically classified into four categories: articulation disorders, fluency disorders, neurologically-based disorders, and organic disorders.

Articulation disorders include substitution or omissions of sounds and other phonological errors. The articulation is impaired as a result of delayed development, hearing impairment, or cleft lip/palate. Fluency disorders also called stuttering are disruptions in the normal flow of speech that may yield repetitions of syllables, words or phrases, hesitations, interjections, prolongation, and/or prolongations. It is estimated that stuttering affects about one percent of the general population in the world, and overall males are affected two to five times more often than females [3]. The effects of stuttering on self-concept and social interactions are often overlooked. The neurologically-based disorders are a broad area that includes any disruption in the production of speech and/or the use of language. Common types of these disorders encompass aphasia, apraxia, and dysarthria. Aphasia is characterized by difficulty in difficulty in formulating, expressing, and/or understanding language. Apraxia makes words, and sentences sound jumbled or meaningless. Dysarthria results from paralysis, lack of coordination or weakness of the muscles required for speech. Organic disorders are characterized by loss of voice quality because of inappropriate pitch or loudness. These problems may result from hearing impairment damage to the vocal cords surgery, disease or cleft palate [4, 5].

In this paper we focus on dysarthria and a Sound Substitution Disorder (SSD) belonging to the articulation disorder category. We propose to extend our previous work [6] by integrating in a new pathologic speech synthesis system a grafting technique that aims at enhancing the intelligibility of dysarthric and SSD speech uttered by American and Acadian French speakers, respectively. The purpose of our study is to investigate to what extent automatic speech recognition and speech synthesis systems can be used to the benefit of American dysarthric speakers and Acadian French speakers with SSD. We intend to answer the following questions.

(i) How well can pathologic speech be recognized by an ASR system trained with limited amount of pathologic speech (SSD and dysarthria)?

(ii) Will the recognition results change if we train the ASR by using variable length of analysis frame, particularly in the case of dysarthria, where the utterance duration plays an important role?

(iii) To what extent can a language model help in correcting SSD errors?

(iv) How well can dysarthric speech and SSD be corrected in order to be more intelligible by using appropriate Text-To-Speech (TTS) technology?

(v) Is it possible to objectively evaluate the resynthesized (corrected) signals using a perceptually-based criterion?

To answer these questions we conducted a set of experiments using two databases. The first one is the Nemours database for which we used read speech of four American dysarthric speakers and one nondysarthric (reference) speaker [7]. All speakers read semantically unpredictable sentences. For recognition an HMM phone-based ASR was used. Results of the recognition experiments were presented as word recognition rate. Performance of the ASR was tested by using speaker dependent models. The second database used in our ASR experiments is an Acadian French corpus of pathologic speech that we have previously elaborated. The two databases are also used to design a new speech synthesis system that allows conveying an intelligible message to listeners. The Mel-Frequency cepstral coefficients (MFCCs) are the acoustical parameters used by our systems. The MFCCs are discrete Fourier transform- (DFT-) based parameters originating from studies of the human auditory system and have proven very effective in speech recognition [8]. As reported in [9], the MFCCs have been successfully employed as input features to classify speech disorders by using HMMs. Godino-Llorente and Gomez-Vilda [10] use MFCCs and their derivatives as front-end for a neural network that aims at discriminating normal/abnormal speakers relatively to various voice disorders including glottic cancer. The reported results lead to conclude that short-term MFCC is a good parameterization approach for the detection of voice diseases [10].

2. Characteristics of Dysarthric and Stuttered Speech

2.1. Dysarthria. Dysarthria is a neurologically-based speech disorder affecting millions of people. A dysarthric speaker has much difficulty in communicating. This disorder induces poor or not pronounced phonemes, variable speech amplitude, poor articulation, and so forth. According to Aronson [11], dysarthria covers various speech troubles resulting from neurological disorders. These troubles are linked to the disturbance of brain and nerve stimuli of the muscles involved in the production of speech. As a result, dysarthric speakers suffer from weakness, slowness, and impaired muscle tone during the production of speech. The organs of speech production may be affected to varying degrees. Thus, the reduction of intelligibility is a common disruption to the various forms of dysarthria.

Several authors have classified the types of dysarthria taking into consideration the symptoms of neurological disorders. This classification is based only upon an auditory perceptual evaluation of disturbed speech. All types of dysarthria affect the articulation of consonants, causing the slurring of speech. Vowels may also be distorted in very severe dysarthria. According to the widely used classification of Darley [12], seven kinds of dysarthria are considered.

Spastic Dysarthria. The vocal quality is harsh. The voice of a patient is described as strained or strangled. The fundamental frequency is low, with breaks occurring in some cases. Hypernasality may occur but is usually not important enough to cause nasal emission. Bursts of loudness are sometimes observed. Besides this, an increase in phoneme-to-phoneme transitions, in syllable and word duration, and in voicing of voiceless stops, is noted.

Hyperkinetic Dysarthria. The predominant symptoms are associated with involuntary movement. Vocal quality is the same as of spastic dysarthria. Voice pauses associated with dystonia may occur. Hypernasality is common. This type of dysarthria could lead to a total lack of intelligibility.

Hypokinetic Dysarthria. This is associated with Parkinson's disease. Hoarseness is common in Parkinson's patients. Also, low volume frequently reduces intelligibility. Monopitch and monoloudness often appear. The compulsive repetition of syllables is sometimes present.

Ataxic Dysarthria. According to Duffy [4], this type of dysarthria can affect respiration, phonation, resonance, and articulation. Then, the loudness may vary excessively, and increased effort is evident. Patients tend to place equal and excessive stress on all syllables spoken. This is why Ataxic speech is sometimes described as explosive speech.

Flaccid Dysarthria. This type of dysarthria results from damage to the lower motor neurons involved in speech. Commonly, one vocal fold is paralyzed. Depending on the place of paralysis, the voice will sound harsh and have low volume or it is breathy, and an inspirational stridency may be noted.

Mixed Dysarthria. Characteristics will vary depending on whether the upper or lower motor neurons remain mostly intact. If upper motor neurons are deteriorated, the voice will sound harsh. However, if lower motor neurons are the most affected, the voice will sound breathy.

Unclassified Dysarthria. Here, we find all types that are not covered by the six above categories.

Dysarthria is treated differently depending on its level of severity. Patients with a moderate form of dysarthria can be taught to use strategies that make their speech more intelligible. These persons will be able to continue to use speech as their main mode of communication. Patients whose dysarthria is more severe may have to learn to use alternative forms of communication.

There are different systems for evaluating dysarthria. Darley et al. [12] propose an assessment of dysarthria through an articulation test uttered by the patients. Listeners identify unintelligible and/or mispronounced phonemes. Kent et al. [13] present a method which starts by identifying the reasons for the lack of intelligibility and then adapts the rehabilitation strategies. His test comes in the form of a list of words that the patient pronounces aloud; the auditor has four choices of words to say what he had heard. The lists of choices take into account the phonetic contrasts that can be disrupted. The design of the Nemours dysarthric speech database, used in this paper, is mainly based on the Kent method. An automatic recognition of Dutch dysarthric speech was carried out, and experiments with speaker independent and speaker dependent models were compared. The results confirmed that speaker dependent speech recognition for dysarthric speakers is more suitable [14]. Another research suggests that the variety of dysarthric users may require dramatically different speech recognition systems since the symptoms of dysarthria vary so much from subject to subject. In [15], three categories of audio-only and audiovisual speech recognition algorithms for dysarthric users are developed. These systems include phone-based and whole-word recognizers using HMMs, phonologic-feature-based and whole-word recognizers using support vector machines (SVMs), and hybrid SVM-HMM recognizers. Results did not show a clear superiority for any given system. However, authors state that HMMs are effective in dealing with large-scale word-length variations by some patients, and the SVMs showed some degree of robustness against the reduction and deletion of consonants. Our proposed assistive system is a dysarthric speaker-dependant automatic speech recognition system using HMMs.

2.2. Sound Substitution Disorders. Sound substitution disorders (SSDs) affect the ability to communicate. SSDs belong to the area of articulation disorders that difficulties with the way sounds are formed and strung together. SDDs are also known as phonemic disorders in which some speech phonemes are substituted for other phonemes, for example, "fwee" instead of "free." SSDs refer to the structure of forming the individual sounds in speech. They do not relate to producing or understanding the meaning or content of speech. The speakers incorrectly make a group of sounds, usually substituting earlier developing sounds for later-developing sounds and consistently omitting sounds. The phonological deficit often substitutes t/k and d/g. They frequently leave out the letter "s" so "stand" becomes "tand" and "smoke," "moke." In some cases phonemes may be well articulated but inappropriate for the context as in the cases presented in this paper. SSDs are various. For instance, in some cases phonemes /k/ and /t/ cannot be distinguished, so "call" and "tall" are both pronounced as "tall." This is called *phoneme collapse* [16]. In other cases many sounds may all be represented by one. For example, /d/ might replace /t/, /k/, and /g/. Usually persons with SSDs are able to hear phoneme distinctions in the speech of others, but they are not able to speak them correctly. This is known as the "fis phenomenon." It can be detected at an early age if a speech pathologist says: "Did you say "fis," don't you mean "fish"?" and the patient answers: "No, I didn't say "fis," I said "fis"." Other cases can deal with various ways to pronounce consonants. Some examples are glides and liquids. Glides occur when the articulatory posture changes gradually from consonant to vowel. As a result, the number of error sounds is often greater in the case of SSDs than in other articulation disorders.

Many approaches have been used by speech-language pathologists to reduce the impact of phonemic disorders on the quality of communication [17]. In the minimal pair approach, commonly used to treat moderate phonemic disorders and poor speech intelligibility, words that differ by only one phoneme are chosen for articulation practice using the listening of correct pronunciations [18]. The second widely used method is called the Phonological cycle [19]. It includes auditory overload of phonological targets at the beginning and end of sessions, to teach formation and a series of the sound targets. Recently, an increasing interest has been noticed for adaptive systems that aim at helping persons with articulation disorder by means of computer-aided systems. However, the problem is still far from being resolved. To illustrate these research efforts, we can cite the Ortho-Logo-Paedia (OLP) project, which proposes a method to supplement speech therapy for specific disorders at the articulation level based on an integrated computer-based system together with automatic ASR and distance learning. The key elements of the projects include a real-time audio-visual feedback of a patient's speech according to a therapy protocol, an automatic speech recognition system used to evaluate the speech production of the patient and web services to provide remote experiments and therapy sessions [20]. The Speech Training, Assessment, and Remediation (STAR) system was developed to assist speech and language pathologists in treating children with articulation problems. Performance of an HMM recognizer was compared to perceptual ratings of speech recorded from children who substitute /w/ for /r/. The findings show that the difference in log likelihood between /r/ and /w/ models correlates well with perceptual ratings (averaged by listeners) of utterances containing substitution errors. The system is embedded in a video game involving a spaceship, and the goal is to teach the "aliens" to understand selected words by spoken utterances [21]. Many other laboratory systems used speech recognition for speech training purposes in order to help persons with SSD [22–24].

The adaptive system we propose uses speaker-dependent automatic speech recognition systems and speech synthesis systems designed to improve the intelligibility of speech delivered by dysarthric speakers and those with articulation disorders.

3. Speech Material

3.1. Acadian French Corpus of Pathologic Speech. To assess the performance of the system that we propose to reduce SSD effects, we use an Acadian French corpus of pathologic speech that we have collected throughout the French regions of the New Brunswick Canadian province. Approximately 32.4% of New Brunswick's total population of nearly 730 000 is francophone, and for the most part, these individuals identify themselves as speakers of a dialect known as Acadian French [25]. The linguistic structure of Acadian French differs from other dialects of Canadian French. The participants in the pathologic corpus were 19 speakers (10 women and 9 men) from the three main francophone regions of New Brunswick. The age of the speakers ranges from 14 to 78 years. The text material consists of 212 read sentences. Two "calibration" or "dialect" sentences, which were meant to elicit specific dialect features, were read by all the 19 speakers. The two calibration sentences are given in (1).

(1)a *Je viens de lire dans "l'Acadie Nouvelle"qu'un pêcheur de Caraquet va monter une petite agence de voyage.*

(1)b *C'est le même gars qui, l'année passée, a vendu sa maison à cinq Français d'Europe.*

The remaining 210 sentences were selected from published lists of French sentences, specifically the lists in Combescure and Lennig [26, 27]. These sentences are not representative of particular regional features but rather they correspond to the type of phonetically balanced materials used in coder rating tests or speech synthesis applications where it is important to avoid skew effects due to bad phonetic balance. Typically, these sentences have between 20 and 26 phonemes each. The relative frequencies of occurrence of phonemes across the sentences reflect the distribution of phonemes found in reference corpora of French spoken in theatre productions; for example, /a/, /r/, and schwa are among the most frequent sounds. The words in the corpus are fairly common and are not part of a specialized lexicon. Assignment of sentences to speakers was made randomly. Each speaker read 50 sentences including the two dialect sentences. Thus, the corpus contains 950 sentences. Eight speech disorders are covered by our Acadian French corpus: stuttering, aphasia, dysarthria, sound substitution disorder, Down syndrome, cleft palate and disorder due to hair impairment. As specified, only sound substitution disorders are considered by the present study.

3.2. Nemours Database of American Dysarthric Speakers. The Nemours dysarthric speech database is recorded in Microsoft RIFF format and is composed of wave files sampled with 16-bit resolution at a 16 kHz sampling rate after low-pass filtering at a nominal 7500 Hz cutoff frequency with a 90 dB/Octave filter. Nemours is a collection of 814 short nonsense sentences pronounced by eleven young adult males with dysarthria resulting from either Cerebral Palsy or head trauma. Speakers record 74 sentences with the first 37 sentences randomly generated from the stimulus word list, and the second 37 sentences constructed by swapping the first and second nouns in each of the first 37 sentences. This protocol is used in order to counter-balance the effect of position within the sentence for the nouns.

The database was designed to test the intelligibility of English dysarthric speech according to the same method depicted by Kent et al. in [13]. To investigate this intelligibility, the list of selected words and associated foils was constructed in such a way that each word in the list (e.g., boat) was associated with a number of minimally different foils (e.g., moat, goat). The test words were embedded in short semantically anomalous sentences, with three test words per sentence (e.g., the boat is reaping the time). The structure of sentences is as follows: "***THE*** *noun1* ***IS*** *verb-ing* ***THE*** *noun2.*"

Note that, unlike Kent et al. [13] who used exclusively monosyllabic words, Menendez-Padial et al. [7] in the Nemours test materials included infinitive verbs in which the final consonant of the first syllable of the infinitive could be the phoneme of interest. That is, the /p/ of reaping could be tested with foils such as reading and reeking. Additionally, the database contains two connected-speech paragraphs produced by each of the eleven speakers.

4. Speech-Enabled Systems to Correct Dysarthria and SSD

4.1. Overall System. Figure 1 shows the system we propose to recognize and resynthesize both dysarthric speech and speech affected by SSD. This system is speaker-dependent due to the nature of the speech and the limited amount of data available for training and test. At the recognition level (ASR), the system uses in the case of dysarthric speech a variable Hamming window size for each speaker. The size giving the best recognition rate will be used in the final system. Our interest to frame length is justified by the fact that duration length plays a crucial role in characterizing dysarthria and is specific for each speaker. For speaker with SSD, a regular frame length of 30 milliseconds is used advanced by 10 milliseconds. At the synthesis level (Text-To-Speech), the system introduces a new technique to define variable units, a new concatenating algorithm and a new grafting technique to correct the speaker voice and make it more intelligible for dysarthric speech and SSD. The role of concatenating algorithm consists of joining basic units and producing the desired intelligible speech. The bad units pronounced by the dysarthric speakers are indirectly identified by the ASR system and then need to be corrected.

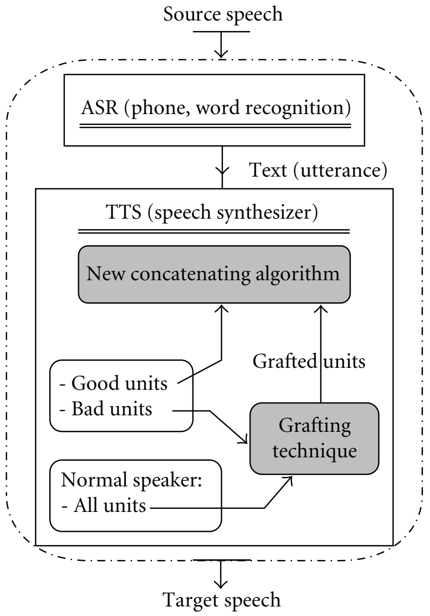
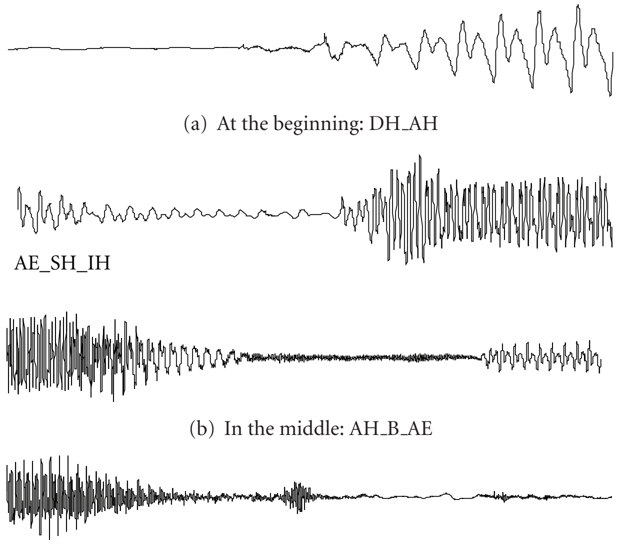


Figure 1: Overall system designed to help both dysarthric speakers and those with SSD.



(a) At the beginning: DH_AH

(b) In the middle: AH_B_AE

(c) At the end: AE_TH

Figure 2: The three different segmented units of the dysarthric speaker BB.

Therefore, to improve them we use a grafting technique that uses the same units from a reference (normal) speaker to correct poorly pronounced units.

4.2. Unit Selection for Speech Synthesis. The communication system is tailored to each speaker and to the particularities of his speech disorder. An efficient alternative communication system must take into account the specificities of each patient. From our point of view it is not realistic to target a speaker independent system that can efficiently tackle the different varieties of speech disorders. Therefore, there is no rule to select the synthesis units. The synthesis units are based on two phonemes or more. Each unit must start and/or finish by a vowel (/a/, /e/ . . . or /i/). They are taken from the speech at the vowel position. We build three different kinds of units according to their position in the utterance.

(i) At the beginning, unit must finish by a vowel preceded by any phoneme.

(ii) In the middle, unit must start and finish by a vowel. Any phoneme can be put between them.

(iii) At the end, unit must start by a vowel followed by any phoneme.

Figure 2 shows examples of these three units. This technique of building units is justified by our objective which consists of facilitating the grafting of poorly pronounced phonemes uttered by dysarthric speakers. This technique is also used to correct the poorly pronounced phonemes of speakers with SSD.

4.3. New Concatenating Algorithm. The units replacing the poorly pronounced units due to SSD or dysarthria are concatenated at the edge starting or ending of vowels (quasiperiodic). Our algorithm always concatenates two periods of the same vowel with different shapes in the time domain. It concatenates /a/ and /a/, /e/ and /e/, and so forth. For ear perception two similar vowels, following each other, sound the same as one vowel, even their shapes are different [28] (e.g., /a/ followed by /a/ sounds as /a/). Then, the concatenating algorithm is as follow.

(i) Take one period from the left unit (LP).

(ii) Take one period from the right unit (RP).

(iii) Use a warping function [29] to convert LP to RP in the frequency domain, for instance, a simple one is $\mathbf{Y} = \mathbf{a}\,\mathbf{X}+\mathbf{b}$. We consider in this conversion the energy and fundamental frequency on both periods. The conversion adds necessary periods between two units to maintain a homogenous energy. Figure 3 shows such a general warping function in the frequency domain.

(iv) Each converted period is followed by an interpolation in the time domain.

(v) The added periods are called step conversion number control. This number is necessary to fix how many conversions and interpolations are necessary between two units.

Figure 4 illustrates our concatenation technique in an example using two units: /ah//b//ae/ and /ae//t//ih/.

4.4. Grafting Technique to Correct SSD and Dysarthric Speech. In order to make dysarthric speech and speech affected by SSD more intelligible, a correction of all units containing those phonemes is necessary. Thus, a grafting technique is used for this purpose. The grafting technique we propose removes all poorly or not pronounced phonemes (silence)

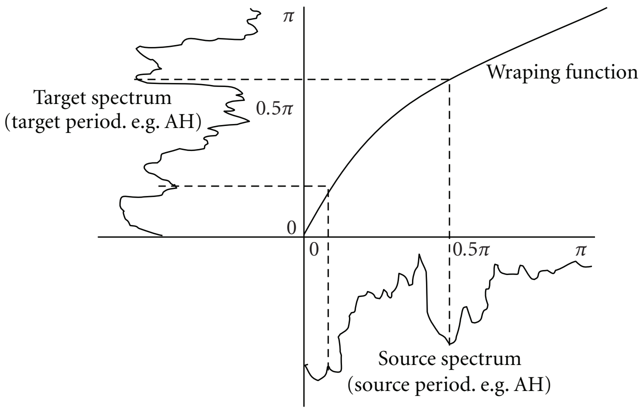


Figure 3: The warping function used in the frequency domain.

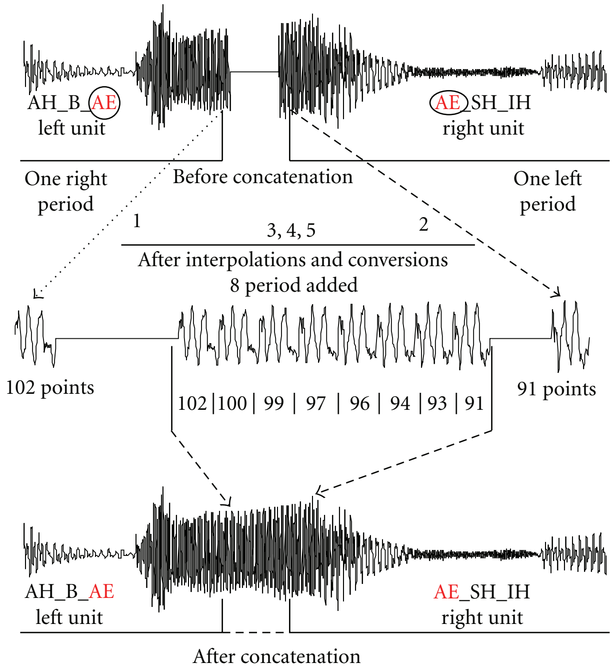
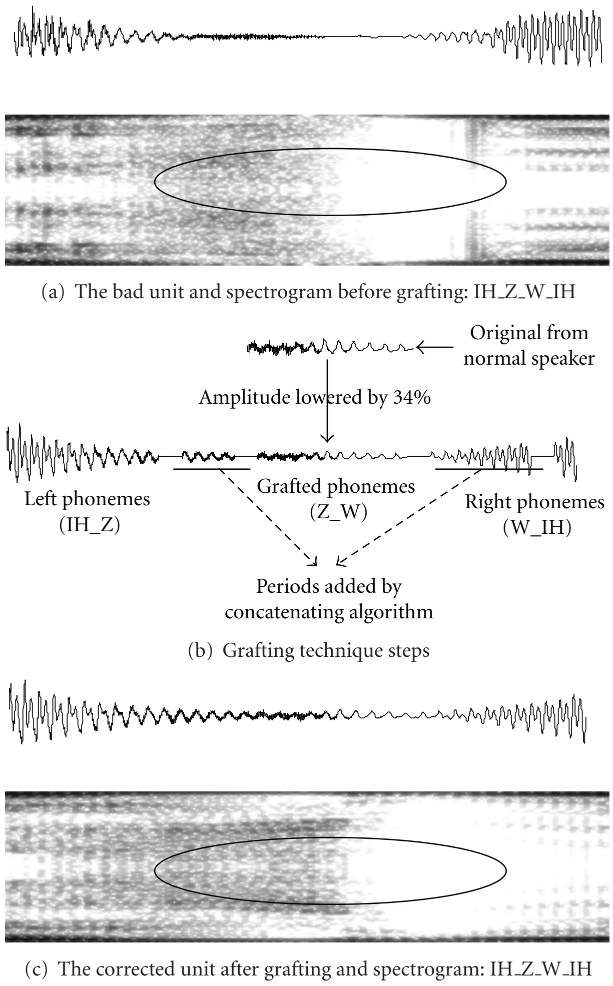


Figure 4: The proposed concatenating algorithm used to link two units: /AH_B_AE/ and /AE_SH_IH/.

following or preceding the vowel from the bad unit, and replaces them with those from the reference speaker. This method has the advantage to provide a synthetic voice that is very close to the one of the speaker. Corrected units are stored in order to be used by the alternative communication system (ASR+TTS). A smoothing at the edges is necessary in order to normalize the energy [29]. Besides this, and in order to dominate the grafted phonemes and hear the speaker with SSD or dysarthria instead of normal speaker, we must lower the amplitude of those phonemes. By iterating this mechanism, we make the energy of unit vowels rising and the grafted phonemes falling. Therefore, the vowel energy on both sides dominates and makes the original voice dominating too. The grafting technique is performed according the following steps.



(a) The bad unit and spectrogram before grafting: IH_Z_W_IH

(b) Grafting technique steps

(c) The corrected unit after grafting and spectrogram: IH_Z_W_IH

Figure 5: Grafting technique example correcting the unit /IH/Z/W/IH/.

1*st step*. Extract the left phonemes of the bad unit (vowel + phoneme) from the speaker with SSD or dysarthria.

2*nd step*. Extract the grafted phonemes of the good unit from the normal speaker.

3*rd step*. Cut the right phonemes of the bad unit (vowel + phoneme) from the speaker with SSD or dysarthria.

4*th step*. Concatenate and smooth the parts above obtained in the three first steps.

5*th step*. Lower the amplitude of signal obtained in step 2, and repeat step 4 till we have a good listening.

Figure 5 illustrates the proposed grafting on an example using the unit /IH/Z/W/IH/ where the /W/ is not pronounced correctly.

4.5. Impact of the Language Model on ASR of Utterances with SSD. The performance of any recognition system depends on many factors, but the size and the perplexity of the vocabulary are among the most critical ones. In our systems, the size of vocabulary is relatively small since it is very difficult to collect huge amounts of pathologic speech.

A language model (LM) is essential for effective speech recognition. In a previous work [30], we have tested the effect of the LM on the automatic recognition of accented speech. The results we obtained showed that the introduction of LM masks numerous pronunciation errors due to foreign accents. This leads us to investigate the impact of LM on errors caused by SSD.

Typically, the LM will restrict the allowed sequences of words in an utterance. It can be expressed by the formula giving the a priori probability, $P(W)$:

$$\begin{aligned} P(W) &= p(w_1,\ldots,w_m) \\ &= p(w_1)\prod_{i=2}^{m} p\left(w_i \mid \underbrace{w_{i-n+1},\ldots,w_{i-1}}_{n-1}\right), \end{aligned} \tag{1}$$

where $W = w_1,\ldots,w_m$ is the sequence of words. In the n-gram approach described by (1), n is typically restricted to $n = 2$ (bigram) or $n = 3$ (trigram).

The language model used in our experiments is a bigram, which mainly depends on the statistical numbers that were generated from the phonetic transcription. All input transcriptions (labels) are fed to a set of unique integers in the range 1 to L, where L is the number of distinct labels. For each adjacent pair of labels i and j, the total number of occurrences $O(i, j)$ is counted. For a given label i, the total number of occurrences is given by

$$O(i) = \sum_{j=1}^{L} O(i, j). \tag{2}$$

For both word and phonetic matrix bigrams, the bigram probability $p(i, j)$ is given by

$$p(i, j) = \begin{cases} \alpha \dfrac{O(i, j)}{O(i)}, & \text{if } O(i) > 0, \\ \dfrac{1}{L}, & \text{if } O(i) = 0, \\ \beta, & \text{otherwise,} \end{cases} \tag{3}$$

where β is a floor probability, and α is chosen to ensure that

$$\sum_{j=1}^{L} p(i, j) = 1. \tag{4}$$

For back-off bigrams, the unigram probablities $p(i)$ are given by

$$p(i) = \begin{cases} \dfrac{O(i)}{O}, & \text{if } O(i) > \gamma, \\ \dfrac{\gamma}{O}, & \text{otherwise,} \end{cases} \tag{5}$$

where γ is unigram floor count, and O is determined as follows:

$$O = \sum_{j=1}^{L} \max[O(i), \gamma]. \tag{6}$$

The backed-off bigram probabilities are given by

$$p(i, j) = \begin{cases} \dfrac{(O(i, j) - D)}{O(i)}, & \text{if } O(i, j) > \theta, \\ b(i)p(j), & \text{otherwise,} \end{cases} \tag{7}$$

where D is a discount, and θ is a bigram count threshold. The discount D is fixed at 0.5. The back-off weight $b(i)$ is calculated to ensure that

$$\sum_{j=1}^{L} p(i, j) = 1. \tag{8}$$

These statistics are generated by using the HLStats function, which is a tool of the HTK toolkit [31]. This function computes the occurrences of all labels in the system and then generates the back-off bigram probabilities based on the phoneme-based dictionary of the corpus. This file counts the probability of the occurrences of every consecutive pairs of labels in all labelled words of our dictionary. A second function of HTK toolkit, HBuild, uses the back-off probabilities file as an input and generates the bigram language model. We expect that the language model through both unigram will correct the nonword utterances. For instance, if at the phonetic level HMMs identify the word “fwee” (instead of “free”), the unigram will exclude this word because it does not exist in the lexicon. When SSD involve realistic words as in the French words “crée” (create) and “clé” (key), errors may occur, but the bigram is expected to reduce them. Another aspect that must be taken into account is the fact that the system is trained only by the speaker with SSD. This yields to the adaptation of the system to the “particularities” of the speaker.

5. Experiments and Results

5.1. Speech Recognition Platform. In order to evaluate the proposed approach, the HTK-based speech recognition system described in [31] has been used throughout all experiments. HTK is an HMM-based speech recognition system. The toolkit was designed to support continuous-density HMMs with any numbers of state and mixture components. It also implements a general parameter-tying mechanism which allows the creation of complex model topologies to suit a variety of speech recognition applications. Each phoneme is represented by a 5-state HMM model with two nonemitting states (1st and 5th state). Mel-Frequency cepstral coefficients (MFCCs) and cepstral pseudoenergy are calculated for all utterances and used as parameters to train and test the system [8]. In our experiments, 12 MFCCs were calculated on a Hamming window advanced by 10 milliseconds each frame. Then, an FFT is performed to calculate a magnitude spectrum for the frame, which is averaged into 20 triangular bins arranged at equal Mel-frequency intervals. Finally, a cosine transform is applied to such data to calculate the 12 MFCCs. Moreover, the normalized log energy is also found, which is added to the 12 MFCCs to form a 13-dimensional (static) vector. This static vector is then expanded to produce a 39-dimensional

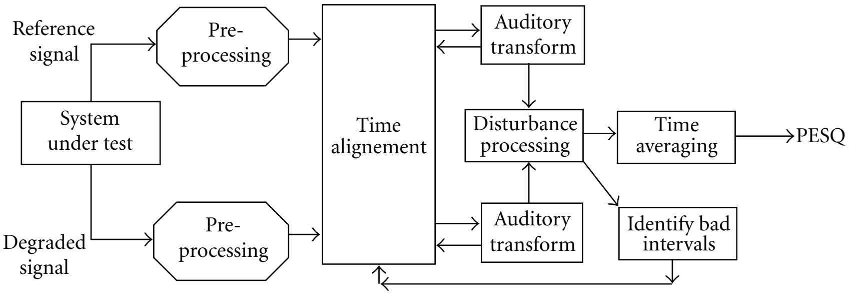


Figure 6: Block diagram of the PESQ measure computation [32].

vector by adding first and second derivatives of the static parameters.

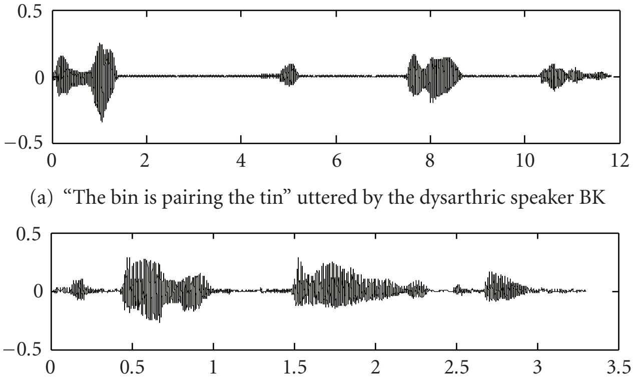
5.2. Perceptual Evaluation of the Speech Quality (PESQ) Measure. To measure the speech quality, one of the reliable methods is the Perceptual Evaluation of Speech Quality (PESQ). This method is standardized in ITU-T recommendation P.862 [33]. PESQ measurement provides an objective and automated method for speech quality assessment. As illustrated in Figure 6, the measure is performed by using an algorithm comparing a reference speech sample to the speech sample processed by a system. Theoretically, the results can be mapped to relevant mean opinion scores (MOSs) based on the degradation of the sample [34]. The PESQ algorithm is designed to predict subjective opinion scores of a degraded speech sample. PESQ returns a score from 0.5 to 4.5, with higher scores indicating better quality. For our experiments we used the code provided by Loizou in [32]. This technique is generally used to evaluate speech enhancement systems. Usually, the reference signal refers to an original (clean) signal, and the degraded signal refers to the same utterance pronounced by the same speaker as in the original signal but submitted to diverse adverse conditions. The idea comes to use the PESQ algorithm since for the two databases a reference voice is available. In fact, the Nemours waveform directories contain parallel productions from a normal adult male talker who pronounced exactly the same sentences as those uttered by the dysarthric speakers. Reference speakers and sentences are also available for the Acadian French corpus of pathologic speech. These references and sentences are extracted from the RACAD corpus we have built to develop automatic speech recognition systems for the regional varieties of French spoken in the province of New Brunswick, Canada [35]. The sentences of RACAD are the same as those used for recording pathologic speech. These sentences are phonetically balanced, which justifies their use in the Acadian French corpora we have built for both normal speakers and speakers with speech disorders. The PESQ method is used to perceptually compare the original pathologic speech with the speech corrected by our systems. The reference speech is taken from the normal speaker utterances. In the PESQ algorithm, the reference and degraded signals are level-equalized to a standard listening level thanks to the preprocessing stage. The gain of the two signals is not known a priori and may vary considerably. In our case, the reference signal differs from the degraded signal since it is not the same speaker who utters the sentence, and the acoustic conditions also differ. In the original PESQ algorithm, the gains of the reference, degraded and corrected signals are computed based on the root mean square values of band-passed-filtered (350–3250 Hz) speech. The full frequency band is kept in our scaled version of normalized signals. The filter with a response similar to that of a telephone handset, existing in the original PESQ algorithm, is also removed. The PESQ method is used throughout all our experiments to evaluate synthetic speech generated to replace both English dysarthric speech and Acadian French speech affected by SSD. The PESQ has the advantage to be independent of listeners and number of listeners.

5.3. Experiments on Dysarthric Speech. Four dysarthric speakers of the Nemours database are used for the evaluation of ASR. The ASR uses vectors contained in varying Hamming Windows. The training is performed on a limited amount of speaker specific material. A previous study showed that ASR of dysarthric speech is more suitable for low-perplexity tasks [14]. A speaker-dependent ASR is generally more efficient and can reasonably be used in a practical and useful application. For each speaker, the training set is composed of 50 sentences (300 words), and the test is composed of 24 sentences (144 words). The recognition task is carried out within the sentence structure of the Nemours corpus. The models for each speaker are triphone left-right HMMs with Gaussian mixture output densities decoded with the Viterbi algorithm on a lexical-tree structure. Due to the limited amount of training data, for each speaker, we initialize the HMM acoustic parameters of the dependent model randomly with the reference utterances as baseline training.

Figure 7 shows the sentence "*The bin is pairing the tin*" pronounced by the dysarthric speaker referred by his initials, BK, and the nondysarthric (normal) speaker. Note that the signal of the dysarthric speaker is relatively long. This is due to his slow articulation. As for standard speech, to perform the estimation of dysarthric speech parameters, the analysis should be done frame-by-frame and with overlapping. Therefore, we carried out many experiments in order to find the optimal frame size of the acoustical analysis window. The tested lengths of these windows are 15, 20, 25, and 30 milliseconds. The determination of the frame size is not controlled only by the stationarity and ergodicity condition, but also by the information contained in each frame. The choice of analysis frame length is a trade-off between having long enough frames to get reliable estimates (of acoustical parameters), but not too long so that rapid events are averaged out [8]. In our application we propose to update the frame length in order to control the smoothness of the parameter trajectories over time. Table 1 shows the recognition accuracy for different lengths of Hamming window and the best result (in bold) obtained for BB, BK, FB, and MH speakers. These results show that the recognition accuracy can increase by 6% when the window length is doubled (15 milliseconds to 30 milliseconds). This leads us to conclude that, in the context of dysarthric speech recognition, the frame length plays a crucial role. The average recognition rate for the four dysarthric speakers is about 70%, which is a very satisfactory result. In order to give an idea about the suitability of ASR for dysarthric speaker assistance, 10 human listeners who have never heard the recordings before are asked to recognize the same dysarthric utterances as those presented to the ASR system. Less than 20% of correct recognition rate has been obtained. Note that in a perspective of a complete communication system, the ASR is coupled with speech synthesis that uses a voice that is very close to the one of the patient thanks to the grafting technique.

The PESQ-based objective test is used to evaluate the Text-To-Speech system that aimed at correcting the dysarthric speech. Thirteen sentences generated by the TTS, for each dysarthric speaker, are evaluated. These sentences have the same structure as those of the Nemours database (*THE noun1 IS verb-ing THE noun2*). We used the combination of 74 words and 34 verbs in "*ing*" form to generate utterances as pronounced by each dysarthric speaker in Nemours database. We also generate random utterances that have never been pronounced. The advantage of using PESQ for evaluation is that it generates an output Mean Opinion Score (MOS) that is a prediction of the perceived quality that would be assigned to the test signal by auditors in a subjective listening test [33, 34]. PESQ determines the audible difference between the reference and dysarthric signals. The PESQ value of the original dysarthric signal is computed and compared to the PESQ of the signal corrected by the grafting technique. The cognitive model used by PESQ computes an objective listening quality MOS ranging between 0.5 and 4.5. In our experiments, the reference signal is the normal utterance which has the code JP prefixed to the filename of dysarthric speaker (e.g., JPBB1.wav), the original test utterance is the dysarthric utterance without correction (e.g., BB1.wav), while the corrected utterance is generated after application of the grafting technique. Note that the designed TTS system can generate sentences that are never pronounced before by the dysarthric speaker thanks to the recorded dictionary of corrected units and the concatenating algorithm. For instance, this TTS system can easily be incorporated in a voicemail system to allow the dysarthric speaker to record messages with its own voice.

The BB and BK dysarthric speakers who are the most severe cases were selected for the test. The speech from the BK speaker who had head Trauma and is quadriplegic was extremely unintelligible. Results of the PESQ evaluation confirm the severity of BK dysarthria when compared with the BB case. Figure 8 shows variations of PESQ for 13 sentences of the two speakers. The BB speaker achieves 2.68 and 3.18 PESQ average for original (without correction) and corrected signals, respectively. The BK speaker affected by the most severe dysarthria achieves 1.66 and 2.2 PESQ average for the 13 original and corrected utterances, respectively. This represents an improvement of, respectively, 20% and 30% of the PESQ of BB and BK speakers. These results confirm the efficacy of the proposed method to improve the intelligibility of dysarthric speech.



(a) "The bin is pairing the tin" uttered by the dysarthric speaker BK

(b) "The bin is pairing the tin" uttered by the normal speaker

Figure 7: Example of utterance extracted from the Nemours database.

5.4. Experiments on Acadian French Pathologic Utterances. We carried out two experiments to test our assistive speech-enabled systems. The first experiment assessed the ASR general performance. The second investigated the impact of a language model on the reduction of errors due to SSD. The ASR was evaluated using data of three speakers, two females and one male, who substitute /k/ by /a/, /s/ by /th/ and /r/ by /a/ and referred to F1, F2, and M1, respectively. Experiments involve a total of 150 sentences (1368 words) among which 60 (547 words) were used for testing. Table 2 presents the overall system accuracies of the two experiments in both word level (using LM) and phoneme level (without using any LM) by considering the same probability of any two sequences of phonemes. Experiments are carried out by using a triphone left-right HMM with Gaussian mixture output densities decoded with the Viterbi algorithm on a lexical-tree structure. The HMMs are initialized with the reference speakers' models. For the considered word units, the overall performance of the system is increased by around 38%, as shown in Table 2. Obviously when the LM is introduced, better accuracy is obtained. When the recognition performance is analyzed at the phonetic level, we were not able to distinguish which errors are corrected by the language model from those that are adapted in the training process. In fact, the use of the speaker-dependent system with LM masks numerous pronunciation errors due to SSD.

Table 1: The ASR accuracy using 13 MFCCs and their first and second derivatives and variable Hamming window size.

Dysarthric	Recognition accuracy (%) for different Hamming window size			
Speaker	15 milliseconds	20 milliseconds	25 milliseconds	30 milliseconds
BB	62.50	63.89	65.28	**68.66**
BK	52.08	55.56	**56.86**	54.17
FB	74.31	76.39	76.39	**80.65**
MH	**74.31**	71.53	70.14	72.92

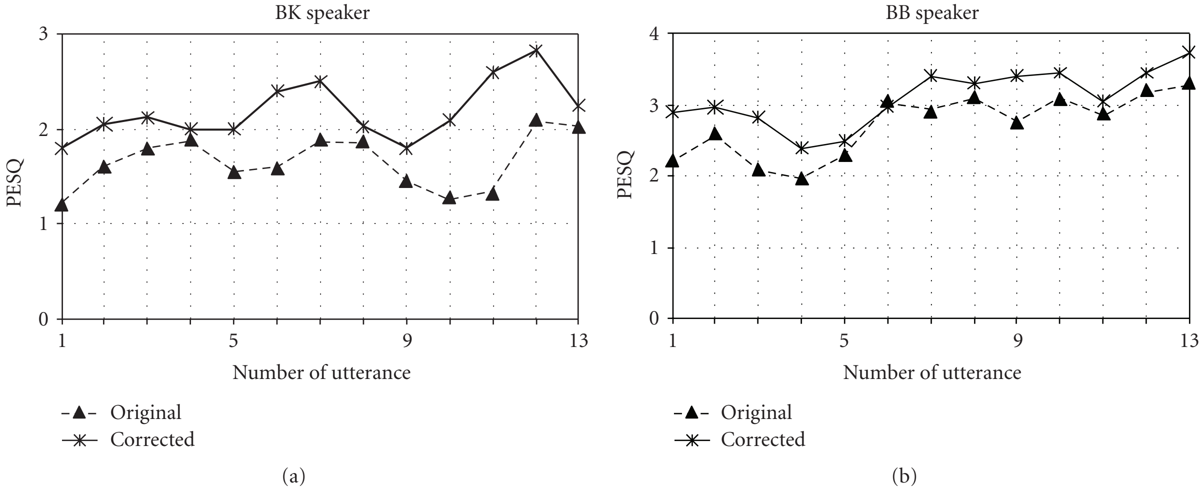


Figure 8: PESQ scores of original (degraded) and corrected utterances pronounced by BK and BB dysarthric speakers.

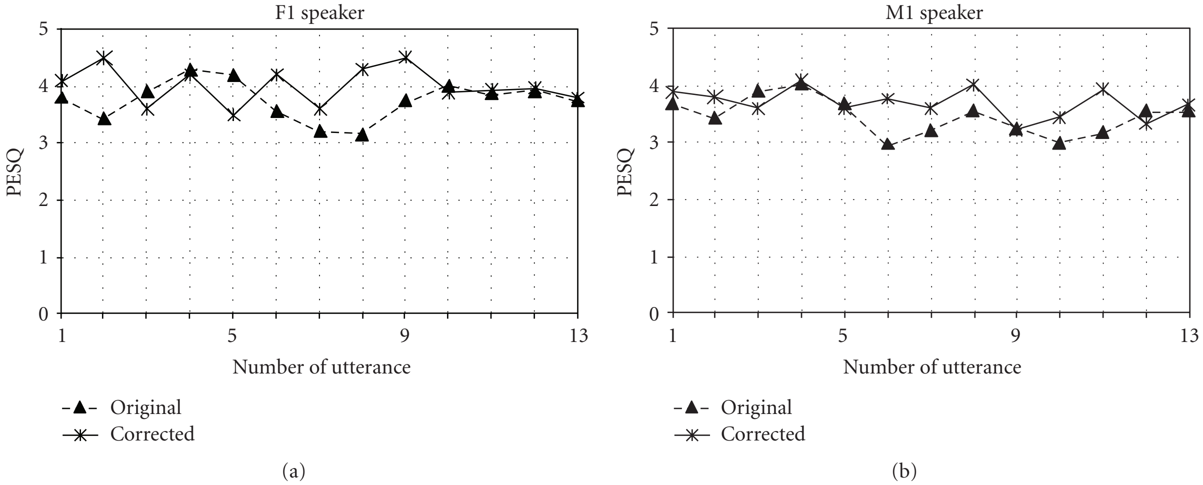


Figure 9: PESQ scores of original (degraded) and corrected utterances pronounced by F1 and M1 Acadian French speakers affected by SSD.

Table 2: Speaker dependent ASR system performance with and without language model and using the Acadian French pathologic corpus.

Speaker	F1 (423/161)	F2 (517/192)	M1 (428/194)	F1 (423/161)	F2 (517/192)	M1 (428/194)
Corr (%)	43.09%	40.87%	46.45%	81.58%	78.44%	83.48%
Del (%)	4.38%	4.96%	4.02%	3.13%	3.26%	2.88%
Sub (%)	52.22%	54.58%	48.47%	15.04%	16.57%	14.55%
	Without bigram-based language model			With bigram-based language model		

The PESQ algorithm is used to objectively evaluate the quality of utterances after correcting the phonemes. The results for F1 who substitutes /k/ by /a/ and M1 who substitutes /r/ by /a/, for thirteen sentences, are given in Figure 9. Even if it is clear that the correction of this substitution disorder is done effectively and is very impressive for listeners, the PESQ criterion does not clearly show this drastic improvement of pronunciation. For speaker F1, 3.76 and 3.98 of PESQ average have been achieved for the thirteen original (degraded) and corrected utterances, respectively. The male speaker M1 achieves 3.47 and 3.64 of PESQ average for the original and corrected utterances, respectively. An improvement of 5% in the PESQ is achieved for each of the two speakers.

6. Conclusion

Millions of people in the world have some type of communication disorder associated with speech, voice, and/or language trouble. The personal and societal costs of these disorders are high. On a personal level, such disorders affect every aspect of daily life. This motivates us to propose a system which combines robust speech recognition and a new speech synthesis technique to assist speakers with severe speech disorders in their verbal communications. In this paper, we report results of experiments on speech disorders. We must underline the fact that very few studies have been carried out in the field of speech-based assistive technologies. We have also noticed the quasiabsence of speech corpora of pathologic speech. Due to the fact that speech pathologies are specific to each speaker, the designed system is speaker-dependant. The results showed that the frame length played a crucial role in the dysarthric speech recognition. The best recognition rate is generally obtained when the Hamming window size is greater than 25 milliseconds. The synthesis system, built for two selected speakers characterized by a severe dysarthria, improved the PESQ by more than 20%. This demonstrates that the grafting technique we proposed considerably improved the intelligibility of these speakers. We have collected data of Acadian French pathologic speech. These data permit us to assess an automatic speech recognition system in the case of SSD. The combination of using both of the language model and the proposed grafting technique has been proven effective to completely remove the SSD errors. We train the systems using MFCCs, but currently we are investigating the impact of using other parameters based on ear modeling, particularly in the case of SSD.

Acknowledgments

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Foundation for Innovation (CFI), and the New Brunswick Innovation Foundation (NBIF) to Sid-Ahmed Selouani (Université de Moncton). The authors would like to thank Mélissa Chiasson and the French Health Authorities of New Brunswick (Beauséjour, Acadie-Bathurst and Restigouche) for their valuable contributions to the development of the Acadian pathologic speech corpus.

References

[1] S. J. Stoeckli, M. Guidicelli, A. Schneider, A. Huber, and S. Schmid, "Quality of life after treatment for early laryngeal carcinoma," *European Archives of Oto-Rhino-Laryngology*, vol. 258, no. 2, pp. 96–99, 2001.

[2] Canadian Association of Speech-Language Pathologists and Audiologists, "General Speech & Hearing Fact Sheet," Report, http://www.caslpa.ca/PDF/fact%20sheets/speechhearingfactsheet.pdf.

[3] E. Yairi, N. Ambrose, and N. Cox, "Genetics of stuttering: a critical review," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 4, pp. 771–784, 1996.

[4] J. R. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*, Mosby, St. Louis, Mo, USA, 1995.

[5] R. D. Kent, *The Mit Encyclopedia of Communication Disorders*, MIT Press, Cambridge, Mass, USA, 2003.

[6] M. S. Yakcoub, S.-A. Selouani, and D. O'Shaughnessy, "Speech assistive technology to improve the interaction of dysarthric speakers with machines," in *Proceedings of the 3rd IEEE International Symposium on Communications, Control, and Signal Processing (ISCCSP '08)*, pp. 1150–1154, Malta, March 2008.

[7] X. Menendez-Padial, et al., "The nemours database of dysarthric speech," in *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP '96)*, Philadelphia, Pa, USA, October 1996.

[8] D. O'Shaughnessy, *Speech Communications: Human and Machine*, IEEE Press, New York, NY, USA, 2nd edition, 2000.

[9] M. Wiśniewski, W. Kuniszyk-Jóźkowiak, E. Smołka, and W. Suszyński, "Automatic detection of disorders in a continuous speech with the hidden Markov models approach," in *Computer Recognition Systems 2*, vol. 45 of *Advances in Soft Computing*, pp. 445–453, Springer, Berlin, Germany, 2007.

[10] J. I. Godino-Llorente and P. Gomez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 2, pp. 380–384, 2004.

[11] A. Aronson, *Dysarthria: Differential Diagnosis*, vol. 1, Mentor Seminars, Rochester, Minn, USA, 1993.

[12] F. L. Darley, A. Aronson, and J. R. Brown, *Motor Speech Disorders*, Saunders, Philadelphia, Pa, USA, 1975.

[13] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek, "Toward phonetic intelligibility testing in dysarthria," *Journal of Speech and Hearing Disorders*, vol. 54, no. 4, pp. 482–499, 1989.

[14] E. Sanders, M. Ruiter, L. Beijer, and H. Strik, "Automatic recognition of Dutch dysarthric speech: a pilot study," in *Proceedings of the 7th International Conference on Speech and Language Processing (ICSLP '02)*, pp. 661–664, Denver, Colo, USA, September 2002.

[15] M. Hasegawa-Johnson, J. Gunderson, A. Perlman, and T. Huang, "HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, vol. 3, pp. 1060–1063, Toulouse, France, May 2006.

[16] W. A. Lynn, "Clinical perspectives on speech sound disorders," *Topics in Language Disorders*, vol. 25, no. 3, pp. 231–242, 2005.

[17] B. Hodson and D. Paden, *Targeting intelligible speech: a phonological approach to remediation*, PRO-ED, Austin, Tex, USA, 2nd edition, 1991.

[18] J. A. Gierut, "Differential learning of phonological oppositions," *Journal of Speech and Hearing Research*, vol. 33, no. 3, pp. 540–549, 1990.

[19] J. A. Gierut, "Complexity in phonological treatment: clinical factors," *Language, Speech, and Hearing Services in Schools*, vol. 32, no. 4, pp. 229–241, 2001.

[20] A.-M. Öster, D. House, A. Hatzis, and P. Green, "Testing a new method for training fricatives using visual maps in the Ortho-Logo-Paedia project (OLP)," in *Proceedings of the Annual Swedish Phonetics Meeting*, vol. 9, pp. 89–92, Lövånger, Sweden, 2003.

[21] H. Timothy Bunnell, D. M. Yarrington, and J. B. Polikoff, "STAR: articulation training for young children," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '00)*, vol. 4, pp. 85–88, Beijing, China, October 2000.

[22] A. Hatzis, P. Green, J. Carmichael, et al., "An integrated toolkit deploying speech technology for computer based speech training with application to dysarthric speakers," in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech '03)*, Geneva, Switzerland, September 2003.

[23] S. Rvachew and M. Nowak, "The effect of target-selection strategy on phonological learning," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 3, pp. 610–623, 2001.

[24] M. Hawley, P. Enderby, P. Green, et al., "STARDUST: speech training and recognition for dysarthric users of assistive technology," in *Proceedings of the 7th European Conference for the Advancement of Assistive Technology in Europe*, Dublin, Ireland, 2003.

[25] Statistics Canada, *New Brunswick* (table), Community Profiles 2006 Census, Statistics Canada Catalogue no. 92-591-XWE, Ottawa, Canada, March 2007.

[26] P. Combescure, "20 listes de dix phrases phonétiquement équilibrées," *Revue d'Acoustique*, vol. 56, pp. 34–38, 1981.

[27] M. Lennig, "3 listes de 10 phrases françaises phonétiquement équilibrées," *Revue d'Acoustique*, vol. 56, pp. 39–42, 1981.

[28] J. P. Cabral and L. C. Oliveira, "Pitch-synchronous time-scaling for prosodic and voice quality transformations," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, pp. 1137–1140, Lisbon, Portugal, 2005.

[29] S. David and B. Antonio, "Frequency domain vs. time domain VTLN," in *Proceedings of the Signal Theory and Communications*, Universitat Politecnica de Catalunya (UPC), Spain, 2005.

[30] Y. A. Alotaibi, S.-A. Selouani, and D. O'Shaughnessy, "Experiments on automatic recognition of nonnative arabic speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2008, Article ID 679831, 9 pages, 2008.

[31] Cambridge University Speech Group, *The HTK Book (Version 3.3)*, Cambridge University, Engineering Department, Cambridge, UK.

[32] P. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton, Fla, USA, 2007.

[33] ITU, "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T Recommendation 862, 2000.

[34] ITU-T Recommendation P.800, "Methods for Subjective Determination of Speech Quality," International Telecommunication Union, Geneva, Switzerland, 2003.

[35] W. Cichocki, S.-A. Selouani, and L. Beaulieu, "The RACAD speech corpus of New Brunswick Acadian French: design and applications," *Canadian Acoustics*, vol. 36, no. 4, pp. 3–10, 2008.