Speech-Based Human and Service Robot Interaction: An Application for Mexican Dysarthric People

Regular Paper

Santiago Omar Caballero Morales1,*, Gladys Bonilla Enríquez1 and Felipe Trujillo Romero1

1 Technological University of the Mixteca, Huajuapan de León, Oaxaca, México
* Corresponding author E-mail: scaballero@mixteco.utm.mx

Received 22 Jul 2012; Accepted 2 Oct 2012

© 2013 Caballero-Morales et al.; licensee InTech. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract Dysarthria is a motor speech disorder due to weakness or poor coordination of the speech muscles. This condition can be caused by a stroke, traumatic brain injury, or by a degenerative neurological disease. Commonly, people with this disorder also have muscular dystrophy, which restricts their use of switches or keyboards for communication or control of assistive devices (i.e., an electric wheelchair or a service robot). In this case, speech recognition is an attractive alternative for interaction and control of service robots, despite the difficulty of achieving robust recognition performance. In this paper we present a speech recognition system for human and service robot interaction for Mexican Spanish dysarthric speakers. The core of the system consisted of a Speaker Adaptive (SA) recognition system trained with normal-speech. Features such as on-line control of the language model perplexity and the adding of vocabulary, contribute to high recognition performance. Others, such as assessment and text-to-speech (TTS) synthesis, contribute to a more complete interaction with a service robot. Live tests were performed with two mild dysarthric speakers, achieving recognition accuracies of 90-95% for spontaneous speech and 95-100% of accomplished simulated service robot tasks.

Keywords Dysarthria, Assistive Technology, Service Robot Control, Speech Recognition, Mexican Spanish

1. Introduction

Service robots are used nowadays for different tasks focused towards the assistance of human beings, especially for elderly people or people with physical disabilities. In Figure 1 some robots used for this purpose are shown: (a) RI-MAN (Robot Interacting with HuMAN), the first robot designed to lift and carry human beings. It has sensors to see, hear and smell, and is used for the purpose of caregiving to elderly people. It can perform tasks such as locating people who are calling out to it, serve them their food, respond to spoken commands and check their sanitary conditions [1]; (b) RIBA (Robot for Interactive Body Assistance), an update of RI-MAN for the same purposes of nursing-care tasks; (c) HAR (Home Assistant Robot), a robot for performing housework tasks such as washing, cleaning dishes and moving furniture; (d) REEM, a service robot which incorporates autonomous navigation and voice and face recognition. It has been used as an assistant to provide
information and guiding in hospitals, hotels and shopping malls [2]. Cosero (Cognitive Service Robot), a robot for domestic service tasks. It makes use of voice and image recognition systems for interaction with users [3].

Figure 1. Examples of service robots used for assistive tasks

In 2010, approximately 5.1% of the total Mexican population had some kind of disability and within this percentage, 50.3% had a motor disability [4]. Assistive robotics can provide significant benefits to these people.

However, in order for these robots to perform accurately the required assistive tasks, they must first understand correctly the command given by the human user. While this can be accomplished with robust speech and image recognition systems, natural language processing and artificial intelligence, interaction may be seriously affected if the person has a physical disability which limits his/her capacity to provide visual (signs) or acoustic (speech) information to the robot. People with muscle atrophy and speech disorders can find difficult to use these robots for their needs, as is the case of people with dysarthria.

Within this context, we focus on the speech aspect of the Human-Robot Interaction (HRI) [5] for Mexican people with dysarthric speech. Hence, in this paper, we present the development of an interface for HRI that makes use of Automatic Speech Recognition (ASR) for communication and service robot control. This interface integrates the following functions in order to accomplish high performance for HRI:

- Dynamically adapt the ASR system for different users.
- Add vocabulary according to the requirements of the user (or his/her therapist) and control the ASR's language model restrictions.
- Reproduce the recognized speech in a more intelligible voice (TTS synthesis)
- Assess the user’s pronunciation patterns to enable robot/human assisted diagnosis and therapeutic tasks.

The ASR interface achieved recognition accuracies of 90-95% for two diagnosed dysarthric speakers and 95-100% for simulated service robot tasks. In this paper we present the details of our findings and the development of the complete interface as follows. In Section 2 a background about dysarthria and related work on ASR is presented, in Section 3 the development of the ASR interface is presented, in Section 4 information about the interface’s tests is presented (e.g., speaker's profile, test vocabulary, comparison of results) and finally, in Section 5 we discuss on our findings and future work.

2. Dysarthria

Dysarthria is a motor speech disability that is often associated with irregular phonation and amplitude, incoordination and restricted movement of speech articulators. Thus, dysarthria includes motor dysfunction of respiration, phonation, resonance, articulation and prosody [6]. As a result of these dysfunctions, dysarthric speech is characterized by the following specific symptoms in the pronunciation of phonemes in the Spanish language [7]:

- Substitution: a phoneme is replaced by another (e.g., /d/ or /g/ is uttered instead of /r/, or /t/ instead of /k/).
- Deletion: a phoneme is omitted (e.g., “niño” is uttered instead of “niño” (kid)), or the whole syllable where the phoneme is present is omitted (e.g., “loj” is uttered instead of “reloj” (watch)).
- Insertion: a phoneme that does not match the spoken phonemes is inserted to support the pronunciation of a phoneme, which is difficult to articulate (e.g., “Enrique” is uttered instead of the name “Enrique”).
- Distortion: a sound that doesn’t match a phoneme pronunciation is uttered as a possible pronunciation for a phoneme with articulation difficulties.

Dysarthria can be caused by a stroke or injury that affects the central nervous system, or by neuronal degenerative diseases like Multiple Sclerosis or Parkinson’s disease [8]. Hence, besides speech disability, people with dysarthria can also be affected by paralysis (or restricted movement) on one side (or both sides) of their bodies, impacting their ability to communicate verbally or manually. Because the type and severity of dysarthria depends on the part of the nervous system that is affected, there are a wide variety of abnormalities among dysarthric people.
2.1 Applications of ASR on Dysarthric Speech

There has been research for the development of technological tools to support people with this disability, especially in the field of ASR research. The use of commercial systems, such as Dragon Naturally Speaking, Microsoft Dictation, VoicePad Platinum and Infovox RA [9, 10, 11, 12], has shown varying levels of recognition (in the range of 50% to 95%) for users with different levels of dysarthria, obtaining the best performance for small vocabularies (10 - 78 words).

Research projects have been developed to improve these systems. In [13] the use of Artificial Neural Networks (ANNs) was explored, which performed better than the commercial system IntroVoice. Significant performances were also obtained in [14] and [15] with Hidden Markov Models (HMMs) for Dutch and Japanese speaking users. In [16] accuracy rates of HMM-based speech recognition of 86.9% were obtained for British speakers with severe dysarthria and a vocabulary of 7-10 words to control electronic devices (radio, TV, etc.). In [17], a system to activate different parts of a web browser was developed with a vocabulary of 47 pre-selected words. However, work is limited for Mexican Spanish speakers.

In Mexico there is knowledge of the computer game “Gravedad” [18], which was developed at the Autonomous University of Yucatan. This game used HMM-based ASR to stimulate children with dyslalia, a speech disorder, to interact with different game scenarios and characters. In [19] a phoneme processing system for the rehabilitation of people with speech disorders using machine learning techniques was developed. This system consisted of modules that allowed the therapist to manage registration and therapy activities for patients (e.g., pronunciation exercises, audiovisual information). These functions were similar to those implemented by the system STRAPTK [20] for British speakers. Note however that the systems reported in [18] and [19] did not address the problem of dysarthric speech, although [20] did it.

3. Development of the ASR Interface for HRI

In Figure 2 the main modules of the ASR interface are shown. The core of the interface is a Speaker – Adaptive (SA) baseline system (see Section 3.1), which is managed by each of the modules of the interface. The first and second modules are proposed for human and robot interaction. In this case, the speaker can use the interface to communicate with other people or with the service robot. This can be accomplished by improving the recognition of the user’s speech and its intelligibility. The third module is proposed to support assessment tasks of dysarthria, either assisted by humans or service robots. In the following sections it is explained how these goals were accomplished.

Figure 2. Main modules of the proposed ASR interface for HRI for dysarthric speakers

3.1 Base ASR System

In Figure 3 the main elements of the base ASR system are shown. To build a robust ASR system, a large training speech corpus is usually needed. Commercial systems are trained with thousands of speech samples from speakers of different ages and gender. These speech corpora are expensive and require a long time to produce, as each speech sample must be labelled at the phonetic and orthographic levels. In the case of Mexican Spanish speech, the corpora are very limited. Additionally, obtaining large quantities of speech samples from dysarthric speakers requires more time and effort, given their disability.

Figure 3. Functional elements of the base ASR system

Hence, we explored the use of a Speaker-Adaptive (SA) system, which consisted of a base ASR system, trained with the speech of a normal speaker and adapted to the speech patterns of a dysarthric speaker. We considered that by designing a special text stimulus for the production of training (and adaptation) speech samples, a reliable base ASR system could be accomplished. Moreover, we explored the effect of continuous (dynamic) adaptation, language model perplexity and control of language model restriction, to improve ASR performance and achieve accuracies comparable to those obtained by commercial systems.

3.1.1 Training Speech Corpus

To develop the base ASR system with limited resources, it was assumed that reliable supervised training could be
accomplished if there were enough speech samples (of all phonemes in the language) in the training speech corpus for acoustic modelling (even if only a single speaker were used as the speech source). For this work, the phonemes defined by the Mexbet alphabet for the Mexican Spanish language were used [21]. An updated version of the alphabet, proposed by the Master in Hispanic Linguistics, Javier Octavio Cuétara [22], is shown in Table 1.

| Description                        | IPA | RFE | Mexbet |
|-----------------------------------|-----|-----|--------|
| Voiceless bilabial stop           | p   | p   | p      |
| Voiceless dental stop             | t   | t   | t      |
| Voiceless velar stop              | k   | k   | k      |
| Voiced bilabial stop              | b   | b   | b      |
| Voiced dental stop                | d   | d   | d      |
| Voiced velar stop                 | g   | g   | g      |
| Voiceless palatal affricate       | ʃ   | ʃ   | ʃ      |
| Voiceless labiodental fricative   | f   | f   | f      |
| Voiced alveolar sibilant          | s   | s   | s      |
| Voiced velar fricative            | x   | x   | x      |
| Voiced palatal fricative          | ʒ   | Z   |        |
| Bilabial nasal                    | m   | m   | m      |
| Palatal nasal                     | ɲ   | ɲ   | ɲ      |
| Alveolar nasal                    | n   | n   | n      |
| Alveolar lateral                  | l   | l   | l      |
| Alveolar trill                    | r   | r   | r      |
| Alveolar flap                     | r   | r   | r      |
| Close front unrounded vowel       | i   | i   | i      |
| Close-mid front unrounded vowel   | e   | e   | e      |
| Open front unrounded vowel        | a   | a   | a      |
| Close-mid back rounded vowel      | o   | o   | o      |
| Close back rounded vowel          | u   | u   | u      |

Table 1. IPA, REF and Mexbet representation of the Mexican Spanish phonemes [22]

Cuétara also proposed the inclusion of the archiphonemes /_DI/, /_GI/, /_N/ and /_/R/ in Mexbet to define the neutralization of the following couples of phonemes: /d/-/t/, /g/-/k/, /n/-/m/, and /s/-/ʃ/ [22]. To represent the pronunciation of the sequence of phonemes /k/ and /ʃ/ (as in “extra”) and the silence, the phonemes /ks/ and /sil/ were added. This led to a final alphabet of 28 phonemes for this work. TranscribEMex [23] was built with this alphabet to phonetically transcribe Mexican Spanish words, which is important for speech data labelling for supervised training of acoustic models.

In Figure 4 the steps followed to obtain the training speech corpus are shown. A representative text (which contained all the Mexican Spanish phonemes) for the training corpus was obtained from the following sources:

- A fragment of the story “Party at the Mountain” [24], which was phonetically balanced and consisted of 102 different words.
- 16 designed phonetically balanced sentences. For new users, this text was considered as a stimulus to obtain speech adaptation data (see Figure 9).

![Figure 4. Steps to obtain the speech corpus for supervised training of the base ASR system](image)

In total, the representative text for the corpus consisted of 205 different words. The frequency of phonemes in the representative text is shown in Figure 5. In [25], for command recognition of speakers with severe dysarthria a minimum of six samples of a word was found to be enough to get accuracies of up to 100%. Based on this, our representative text was considered to be well balanced to provide enough speech samples of a phoneme (six being the minimum number of occurrences of a phoneme). In Figure 6 the frequency of phonemes of the sub-set of the representative text corresponding to the adaptation stimuli (16 sentences) is shown.

To obtain the speech data for the corpus the representative text was used as a stimulus. This text was read five times by a person (male) with normal speech, who had the standard pronunciation of the centre region of Mexico. The speech was recorded with a Sony ldx800 recorder with a sampling frequency of 16 kHz monaural in WAV format. Then this data was labelled manually at the word (orthographic) and phonetic level with WaveSurfer. Then the speech data was coded into Mel Frequency Cepstral Coefficients (MFCCs). The front-end used 12 MFCCs plus energy, delta and acceleration coefficients [26]. With the realization of this step the training speech corpus was finished and ready for the training of the base ASR system.
3.1.2 Acoustic Models

The acoustic models are the pattern recognition core of the ASR system and are initialized and re-estimated with the data of the training speech corpus. The technique used for acoustic modelling was a Hidden Markov Model (HMM) and the implementation tool was HTK [26]. An HMM was built for each of the 28 phonemes in the Mexican language (see Figure 5). These HMMs had standard three-state, left-to-right architecture [26].

A continuous probability distribution, which models the observation probability of a given acoustic signal, is associated to each state of an HMM [27]. These observation probabilities are modelled as a mixture of Gaussian distributions, being the performance of the ASR system dependent of the number of mixtures [26]. For recognition performance of dysarthric speech with limited training data, this was considered a main factor.

3.1.3 Language Model and Lexicon

The Language Model (LM) represents a set of rules or probabilities that restrict the recognized sequence of words from the ASR system to valid sequences (e.g., guides the search algorithm to find the most likely sequence of words that best represents an input speech signal). Commonly, N-grams are used for the LM and for this work bigrams (N=2) were used for continuous speech recognition [26, 28].

The Lexicon, which specifies the sequences of phonemes that form each word in the application's vocabulary, was built with the TrancribEMex tool.

3.1.4 The Search Algorithm

The Viterbi algorithm is widely used for speech recognition [28]. This task consists of finding (searching) the sequence of words that best matches the speech signal. Viterbi decoding was implemented with HTK, which requires the following elements: (1) MFCC-coded speech to be recognized, (2) acoustic models, (3) language model, (4) lexicon, (5) scale grammar factor and (6) list of phonemes. The interface manages the construction of each one of these elements and the execution of the Viterbi algorithm to recognize speech. More details are presented in Section 3.2.

3.1.5 Control Variables

- **Gaussian Mixtures**: because the performance of the ASR is dependent of the number of Gaussian mixtures used for acoustic modelling, this was considered as the first variable to be manipulated by the interface's user.
- **LM Perplexity**: Two metrics are commonly used to measure the performance of a LM: (1) Word Error Rate (WER) and (2) perplexity. WER is dependent of the ASR system, as it is estimated from the word output sequences generated by the ASR system. In some cases, low WER correlates with low perplexity [14]. For dysarthric speech recognition, low perplexity is recommended to deal with the effect of slow articulation of speech [14]. Perplexity is not ASR dependent and thus, can be estimated faster than WER [29]. Perplexity increases as the vocabulary grows in size, but the use of an N-gram LM can reduce perplexity for large vocabularies, as it restricts...
the possible sequences of words to the most likely sequences. However, to accomplish this, the test vocabulary must be known in advance by the ASR system [29]. To deal with this issue, we considered building the LM on-line while using the interface, thus constantly updating the LM to allow advanced knowledge of the test vocabulary to reduce perplexity. Hence, updating the vocabulary and the LM were considered as the second variable to be manipulated by the user.

In addition, a third variable was considered, the LM’s scale grammar factor. This factor is the amount by which the LM probability is scaled before being added to each token as it transits from the end of one word to the start of the next [26]. As this factor increases, the recognizer relies more on the LM instead of the acoustic signal to predict what the speaker said (e.g., the LM restrictions have more importance). Hence, the scale grammar factor can be used to reduce the perplexity of the LM during speech recognition.

Thus, to accomplish control of the LM’s perplexity, the following functions in the interface were implemented: (1) adding of vocabulary; (2) manipulation of scale grammar factor; and (3) cumulative estimation of bigrams (LM) considering each word, or sequence of words, added to the system.

3.2 The Speech Interface

The speech interface was programmed in Matlab© 2008 with the GUIDE toolkit. In Figure 7 the starting window of the interface is shown. By pressing each button the user gets access to each of the modules of the interface, which perform the tasks introduced in Figure 2.

![Figure 7. Starting window of the speech interface](image)

3.2.1 Build the Speech Recognizer

In Figure 8 the first module of the interface is shown, which is used for building the base ASR system. Here the user can manipulate the first control variable by establishing the number of Gaussian components required to build the acoustic models of the ASR system.

The user can set the number of Gaussians in “No. Gaussians (1-8)” and then, by pressing the button “Construye Reconocedor” (Build Recognizer), the ASR’s HMMs are automatically built and trained with the stored training speech corpus. On this data the resulting HMMs are evaluated, displaying the word recognition rate (% Corr) and the word recognition accuracy (% Acc, see Eq.2). As shown in Table 2, the performance is higher when the number of mixture Gaussian components is increased. A maximum of eight components was considered.

![Figure 8. Module to Build the Speech Recognizer](image)

| No. Gaussian Mixtures | % Acc | No. Gaussian Mixtures | % Acc |
|-----------------------|-------|-----------------------|-------|
| 1                     | 93.02 | 5                     | 97.81 |
| 2                     | 94.92 | 6                     | 98.45 |
| 3                     | 97.39 | 7                     | 97.88 |
| 4                     | 97.25 | 8                     | 97.81 |

Table 2. Accuracy (Acc) of the base ASR system built with variable number of Gaussian mixtures for HMM modelling

When the base ASR system is built, the user can access the second module, which is explained in the next section.

3.2.2 Adapt the Speech Recognizer

In Figure 9 the second module of the interface is shown, which is used to adapt the base ASR system to the dysarthric user’s voice. Commercial ASR systems are trained with thousands or millions of speech samples from different speakers. When a new user wants to use such a system, it is common to ask the user to read some words or texts (stimuli) to provide speech samples that are then used by the system to adapt its acoustic models to the pattern of the new user’s voice. For this work, Maximum Likelihood Linear Regression (MLLR) [30] was the adaptation technique used to make the ASR system usable for other speakers. MLLR is based on the assumption that a set of linear transformations can be applied to the parameters of the Gaussian components of the ASR system’s HMMs (in this case, mean and variance) to reduce the mismatch between these HMMs and the adaptation data. A regression class tree with 32 terminal nodes was used for the dynamic implementation of MLLR adaptation [26].

If the user is new, he/she must write his/her name in the field “Escriba Nombre de Usuario” (Write User’s Name).
When the user finishes this task and presses “Enter” the interface automatically saves the name and updates the list of registered users, which is shown in the pop-up menu “Seleccione Usuario” (Select User). When the user selects his/her name from this list/menu, the user’s directories are created (or loaded) for the adaptation (or re-adaptation) of the ASR system. To start adaptation the user must record all 16 sentences shown in the push buttons. These can be recorded in any order. When the user presses a sentence’s button, this starts the recording of the user’s speech. When the user ends reading the stimulus, then he/she must press the sentence button again to end the recording process. Especially for dysarthric speakers, the reading speed can be slow, so enabling the user to record speech with variable length was a priority for the design of the interface.

Note that this kind of adaptation is usually performed once before the new speaker uses the system (e.g., static adaptation). In commercial systems, if the speaker wishes to improve adaptation, he/she needs to read other stimuli. For our system we incorporated this task within the use of the ASR system, so adaptation can be made while performing speech recognition (e.g., dynamic adaptation). This is further explained in the next section.

3.2.3 Speech Recognizer

The third module of the interface, shown in Figure 10, consists of the ASR system for communication of dysarthric speakers. Initially the user must select his/her name in the “Seleccione Usuario” (Select User) pop-up menu. When selecting the user’s name his/her adapted acoustic models are automatically loaded. There is also the button “Crea Modelo de Lenguaje” (Create Language Model), which builds the ASR’s LM by considering the vocabulary words/sentences displayed in the window “Frases Vocabulario” (Vocabulary Sentences). This is an informative list about the vocabulary stored in the system and available to be recognized.

![Figure 9. Module to Adapt the Speech Recognizer](image9.png)

Figure 9. Module to Adapt the Speech Recognizer

There is a button next to each sentence button labelled as “Escuchar” (Listen). This is to allow the user to listen to his/her speech sample in order to verify if it was recorded correctly. Hence, the user can record the adaptation speech as many times as needed. At the end of the recording tasks, the user just needs to press the button “Adaptar” (Adapt). By doing this, the interface automatically creates (or loads) the personalized MLLR directories to create (or re-estimate) the adaptation transformations for that user, parameterizes the speech data and performs the MLLR adaptation. The accuracy results (% Corr, % Acc) (see Eq. 2) of the base ASR system on the adaptation data, before (in “Salida Rec. Base”, Base Recognizer’s Output) and after MLLR adaptation (in “Salida Rec. Adaptado”, Adapted Recognizer’s Output), are shown for comparison purposes.

![Figure 10. Module of the Speech Recognizer](image10.png)

Figure 10. Module of the Speech Recognizer

This module allows the user to add more vocabulary to the list shown in “Frases Vocabulario” and thus reduce perplexity (see Section 3.1.5). The new vocabulary must be typed in the form “Añadir Nuevas Frases o Palabras” (Add New Sentences or Words) in UPPER case format. By immediately pressing the “Crea Modelo de Lenguaje” button, the interface automatically updates the ASR’s lexicon (by managing TranscribEMex) and the LM. Another parameter that can be set is the scale grammar factor (see Section 3.1.5) to increase the influence of the LM in the recognition process. The grammar factor’s value can be set in the form “Valor de Ajuste (1-30)”
(Adjustment Value (1-30)). The range for the grammar factor was set to 1-30 as it was observed that, for dysarthric speakers, maximum recognition accuracy is achieved with values over 20 [31]. For non-disordered speech, usually a value of five is used [26]. In the interface this parameter can be changed at any moment without the need to re-start the system. Thus, ASR performance can be adjusted in real time. To start using the speech recognition function for communication the user must press the button “Ejecuta Reconocedor” (Execute Recognizer). This button starts recording when pressed once and finishes the recording process when pressed again. Internally, when recording finishes, the interface performs parametrization of the recorded speech (wav to mfc conversion), managing the HTK to perform Viterbi decoding (recognition), integrating the updated Lexicon, LM and adapted HMMs. Viterbi is executed to provide unadapted (“Salida Original”) and adapted (“Salida Adaptada”) word outputs for the spoken sentences. Additionally, the speech’s waveform is plotted. The word output of the adapted system is then given to a speech synthesizer, which “reads” these words with a more intelligible voice. For this purpose we accessed to the Windows XP Speech Application Programming Interface (SAPI) version 5.0. The voice used for synthesis was Isabel from ScanSoft for Spanish.

Another function of this module is to allow dynamic adaptation of the user’s HMMs. This procedure showed improvements for severe dysarthric speakers [20]. This was implemented as an additional option to the adding of vocabulary in “Añadir Nuevas Frases o Palabras”. Any text written in that form is a stimulus candidate to be read, recorded and added to the user’s personal adaptation speech library. If the user wishes to use any text for adaptation he/she must press the button “Grabar para Adaptación” (Record for Adaptation) located under the form. Internally, each recording is associated to the stimulus text written in the form and there is no restriction on the words that can be added. When the user (or the therapist) considers that enough samples have been recorded, he/she just needs to press the button “OK” to perform re-adaptation with all the accumulated speech samples from the user. Also, by pressing “OK” the Lexicon and LM are updated.

3.2.4 Phoneme Confusion Patterns - Assessment

Assessment of dysarthric speech is initially based on intelligibility tests applied by a human listener. In these tests, the speaker utters a specific set of words, which are heard by a therapist who identifies (recognizes) abnormalities in the pronunciation of those words. The proposed module follows the same procedure, with the difference being that the listener is the ASR system.

Research has shown that ASR performance is highly correlated to the speaker’s intelligibility [31]. Severe dysarthria leads to low intelligibility and poor ASR performance, while high intelligibility and ASR performance is observed with speakers with mild dysarthria. In Figure 11 human assessment of dysarthric speech is compared to ASR performance. Speech data from ten English speakers from the Nemours Database of Dysarthric Speech [32] was used for this comparison.

The intelligibility of each speaker was measured by the Frenchay Dysarthria Assessment (FDA) test [32], which is based on human perception. Based on this test, speakers BK, BV, RK, RL and SC, who have the lowest scores, are classified as speakers with severe dysarthria (low intelligibility). Speaker JF can be classified with moderate dysarthria and speakers BB, FB, LL and MH with mild dysarthria (high intelligibility).

ASR on the speech data was performed with the following systems:

- BASE, an ASR system trained with the speech data of 92 English speakers of the Wall Street Journal (WSJ) Database [33].
- MLLR, the adapted BASE system.

![Figure 11. Comparison of recognition performance: human assessment (FDA), unadapted (BASE) and adapted (MLLR) ASR systems [31].](image)

As presented, the performance of human assessment and computer recognition followed similar patterns. The correlation coefficient between the FDA scores and the performance of the unadapted (BASE) ASR system was 0.67. On the other hand, the correlation between the FDA scores and the performance of the adapted (MLLR) ASR system was 0.82. Both results were significant at the 1% level, which gives confidence about the recognizers displaying a similar performance trend as humans when exposed to speech with different levels of dysarthria. Thus, ASR performance can be used to estimate a speaker’s level of dysarthria. This work served as a basis for the application developed to support the assessment of Mexican dysarthric speech.

Because high ASR performance depends on accurate recognition of phonemes, significant abnormalities in the
pronunciation of phonemes (see Section 2) would affect ASR performance. Tests, such as the FDA, use specific stimuli words to detect such abnormalities. This information is then used to support the assessment of the speaker’s level of dysarthria. For Mexican speakers, this task is supported by the use of the 49 words mentioned in Section 3.1.1.

In this way, the fourth module of the interface, shown in Figure 12, was developed to provide information about confusion patterns in the articulation of phonemes by the speaker. This information is visually shown in the form of a phoneme confusion matrix, as presented in Figure 13. This can be used by the therapist (or a robot) to detect significant abnormalities in the user's speech to define specific therapy tasks. Also, it can be used to support the assessment of a speaker’s level of dysarthria (see Section 4.1).

In the panel “Ficha de Articulación” (Articulation Data) the user selects his/her name in the pop-up menu “Seleccione Usuario” (Select User). By doing this, the interface loads the adapted acoustic models that belong to the user. Then, the user must read and record the sequences of words shown in each button of the panel (as in Section 3.2.2). The words shown in Figure 12 are the 49 words used by the therapists to assess the level of dysarthria of a Mexican Spanish speaking user (see Section 3.1.1). After all the speech samples are recorded, the user just needs to press the button “Matriz de Confusión Fonética” (Phoneme Confusion Matrix) to estimate the confusion patterns in the assessment data. As shown in Figure 13, the vertical axis “Estimulo - Entrada de Voz” represents the Stimulus - Speech Input, and the horizontal axis “Respuesta - Salida del Reconocedor” represents the Response - Output of the Speech Recognizer.

The reliability of the assessment tool relies on the classification of phonemes. Commonly, classification is performed by perception analysis or by time alignment tools. HTK can estimate phoneme confusion matrices from the alignment of the phoneme reference transcription of the speech (P) and the ASR system's phoneme output (P̂). The use of a standard dynamic programming (DP) tool to align two symbol strings (in this case, phoneme sequences) can lead to unsatisfactory results when a precise alignment is required between P and P̂. This is because these tools commonly use a distance measure which is “0” if a pair of phonemes is the same and “1” otherwise. Although HTK can provide such alignments, its distance measures are empirically derived and a correct match has a score of “0”, an insertion and a deletion receive a score of “7” and a substitution a score of “10” [26]. Hence, although the alignment generated by this tool is more effective that those that use “1” and “0” scores, it can be further improved as discussed in [34], where scores based on acoustic similarity were used. These scores for the Mexican Spanish phonemes were estimated by computing the following empirically-derived equation:

\[
\text{Sim}(p^i, \hat{p}^i) = 5 \cdot P_{\text{ST}}(q^i, \hat{q}^i) - 2
\]  (1)

In Eq. 1, Sim\((p^i, \hat{p}^i)\) is the similarity score matrix for a pair of phonemes \(\{p^i, \hat{p}^i\}\) from the alignment of the sequences P and P̂. In [34] these scores were estimated by re-scaling a normalized speaker-independent confusion-matrix \(P_{\text{ST}}(q^i, \hat{q}^i)\), pooled over the data of 92 British-English speakers, where \(q^i\) and \(\hat{q}^i\) were the respective elements of the phoneme sequences \(\{q, \hat{q}\}\) aligned from the associated data. In this case, \(P_{\text{ST}}(q^i, \hat{q}^i)\) was obtained from the phoneme confusion-matrix generated by HTK for the training speech corpus. Hence, a match received the highest score (+3) if the confusion probability in the \(P_{\text{ST}}\) confusion-matrix was high (i.e., \(\geq 0.95\)), with very low confusions receiving the lowest score (-2) (i.e., \(< 0.05\)). After the similarity scores were computed, these were integrated into a DP algorithm to perform alignment and classification of phonemes. The DP algorithm used for
this work is a variation of the symmetric DTW (Dynamic Time Warping) algorithm presented in [35].

The phoneme sequences \( \hat{p} \) were obtained from the unconstrained execution of the adapted Speech Recognizer Module (see Section 3.2.3). Thus, for this recognition task, a phoneme-bigram LM was used without the phoneme-to-words restrictions determined by the Lexicon. The results are discussed in Section 4.1.

4. Performance Tests

The interface was installed in a netbook PC with Windows XP and the following hardware: 1GB RAM and Intel Atom Processor N570 at 1.66 GHz. The microphone was integrated into a headset connected to the netbook.

Initially the Adaptation and Recognition Modules were tested with two users with normal speech (a female and a male student). The vocabulary for the test consisted of 12 sentences (see Table 3) used for control of a simulated robot platform (BOT) and manipulator (CUBE). The simulation software used for these tests was Roboworks©. The simulation of one control sentence, “CUBE TOMA EL VASO” (CUBE take the glass), is shown in Figure 14. Each user read each sentence ten times, thus, each speaker uttered 120 sentences. Only whole-sentences were considered. The ASR performance is presented in Table 4.

![Simulated execution of the recognized spoken command “CUBE TOMA EL VASO”](image)

**Figure 14.** Simulated execution of the recognized spoken command “CUBE TOMA EL VASO”: normal speech

| No | Command Sentence               |
|----|--------------------------------|
| 1  | BOT AVANZA RAPIDO DOS METROS   |
| 2  | BOT RETROCEDE LENTO            |
| 3  | BOT GIRA CUARENTA Y CINCO GRADOS A LA IZQUIERDA |
| 4  | BOT GIRA CUARENTA Y CINCO GRADOS A LA DERECHA |
| 5  | CUBE SIRVE BOTELLA             |
| 6  | CUBE TOMA EL VASO              |
| 7  | BOT SAL POR PUERTA UNO         |
| 8  | BOT ENTRA POR PUERTA DOS       |
| 9  | BOT SIRVE LA COPA              |
| 10 | BOT AVANZA LENTO DOS METROS    |
| 11 | CUBE INICIO                    |
| 12 | BOT DETENTE                   |

**Table 3.** Control sentences for simulated platform: normal speech

| Test User | Fails/Total | %Correct Sentences |
|-----------|-------------|--------------------|
| Male      | 4/120       | 96.67              |
| Female    | 7/120       | 94.17              |

**Table 4.** Performance of the Speech Interface for control sentences: normal speech

These results gave confidence about the reliability of the system when tested by different speakers from those used to train it. As the recognition of correct sentences was over 94% for both speakers, the word recognition accuracy can be assumed to be significantly higher. For the experiments with dysarthric speech, the metrics used to measure the performance of the system was Word Recognition Accuracy (Acc), which is computed as:

\[
Acc = \frac{N-D-S-I}{N} \times 100\%
\]

where \( D \), \( S \) and \( I \) are deletion, substitution and insertion errors in the recognized speech (text output of the ASR module). \( N \) is the number of words in the correct ASR’s output [26]. In the following section the tests performed with dysarthric speakers are presented.

### 4.1 Experiments with Dysarthric People

The authorities of the local SNDIF centre in the city of Huajuapan de León (Oaxaca) provided the support to search and recruit volunteers to participate in this work. During the search process some requirements were established in accordance to the recommendations of the centre’s therapists. Thus, the basic requirements of possible candidates for this work were defined as:

- retention of cognitive understanding (e.g., dysarthria not caused by progressive neurological disorders which affect learning or cognitive abilities such as Alzheimer’s);
- without diagnosis of a language understanding impairment (e.g., the participant must understand the language, word meanings, etc.);
- over 15 years old (younger participants require special supervision);
- professional assessment of dysarthria.

After a period of two months we got collaboration from two participants that fulfilled these requirements. For confidentiality reasons these participants were identified as GJ and MM. In Table 5 their general clinical profile is shown. For MM, some information is missing because he was contacted by personal references (not by means of the DIF centre) and thus, no formal record of his condition was available. Professional assessment of MM’s dysarthria was performed with speech recordings. Because GJ was almost blind, the interface was operated by a family member.
To test the speech interface, initially GJ and MM had to pass through the Adapt Module before using the Speech Recognizer Module. The Speaker Adaptive (SA) ASR system was tested with different amounts of adaptation data to study the effect of static and dynamic adaptation on ASR performance. Three adaptation conditions were considered, leading to the following configurations:

- **SA ASR System I**: Base ASR system adapted with only the 16 sentences of the Adapt Module (static adaptation).
- **SA ASR System II**: SA ASR System I adapted with 11 additional sentences while using the Speech Recognizer Module (dynamic adaptation I).
- **SA ASR System III**: SA ASR System II adapted with 11 additional sentences while using the Speech Recognizer Module (dynamic adaptation II).

All the sentences used for the experiments were spontaneous sentences related to GJ’s and MM’s activities at their home. These were added to the system’s lexicon and LM prior to the test sessions. Each system’s configuration was tested with 50 sentences with a total of 275 unique words. These sentences were different from those used for static and dynamic adaptation. The Latin-American Spanish version of Dragon NaturallySpeaking © (Ver. 10.00.200.161, 2008) (LTN Dragon) was used for comparison purposes. For this system, the user performed the speaker adaptation task, which consisted of reading one out of ten narratives. In this case, the narrative titled “Adventures of Pinocchio” was selected, which consisted of 310 unique words. Once that the adaptation task was completed, the LTN Dragon was tested in dictation mode with the same 50 spontaneous sentences. The results of the test sessions are presented in Table 6. The performance of the interface, using a grammar scale factor of 20, is compared to the performance of other commercial systems and research projects, including human transcription of normal speech.

As presented, the SA ASR systems had a performance of 93.67%-94.94% for GJ, and 90.04%-94.70% for MM. Non-adapted performance achieved recognition accuracies of around 40%, which is similar to the information shown in Figure 11. Thus, standard MLLR adaptation can improve recognition of dysarthric speech.

The performance of the adapted systems is comparable to human transcription (96%-98%) and commercial ASR for normal speech [36]. When compared with systems adapted (or developed) for dysarthric speakers, this interface achieved performance comparable to those for small vocabularies (<100 words) [9, 16] and a similar level of dysarthria [9]. For a system with a similar test vocabulary (300 words) and a speaker with a similar level of dysarthria [11], the interface achieved a higher performance. The use of Dragon NaturallySpeaking had the following performance: 83.50% for GJ and 82.40% for MM. Hence, the proposed system outperforms both commercial and research systems, when compared and tested with similar speaker’s conditions. These results are significant for human communication purposes.

### Table 5. Clinical profile of the dysarthric speakers GJ and MM.

| Name: GJ | Age: 64 | Gender: Male |
|----------|---------|---------------|
| Pathologies: | mild dysarthria caused by a stroke; left hemiplegia (paralysis of the left arm, leg, and trunk); 90% loss of sight; scoliosis (spine is curved from side to side) |

| Name: MM | Age: 37 | Gender: Male |
|----------|---------|---------------|
| Pathologies: | mild dysarthria caused by a tracheotomy; |

### Table 6. Accuracy of the Speech Interface and comparison with other systems: human and commercial ASR on normal speech [36]; commercial and developed systems on dysarthric speech with intelligibility levels: high [9], moderate [11], and low [16].

| Speech | System | Accuracy |
|--------|--------|----------|
| Normal [36] | Human Recognition | 96% - 98% |
| | Commercial ASR System (≤ 1000 words) | 80% - 96% |
| Dysarthric (foreign language) | Dragon Dictate (Version 1.01A) [9] ("Pledge of Allegiance", 24 words) | 80% - 100% |
| | Dragon Naturally Speaking [11] (≈ 300 words) | 77.28% - 82.20% |
| | STARDUST [16] | 88.5% - 95.4% |

| Dysarthric (Mexican Spanish language) | 275 words |
|-------------------------------------|---------|
| SA ASR System I | GJ 93.67% |
| SA ASR System II | 94.51% |
| SA ASR System III | 94.94% |
| LTN Dragon | 83.50% |
| MM | |
| SA ASR System I | 90.04% |
| SA ASR System II | 93.54% |
| SA ASR System III | 94.70% |
| LTN Dragon | 82.40% |

In order to perform accurately assistive tasks, these must be correctly recognized as whole sentences (as shown in Table 4). For this, 20 test sentences were randomly selected from the original set of 50 to be associated with assistive tasks. These consisted of “actions” related to commanding a service robot to go to a particular point and push an object located in that point to its point of origin or to another point. For these tasks and the nature of the users (dysarthric people), we considered the NAO © robot (N40, H25) as an appropriate choice. To perform the simulation of the tasks we used the following software:

- Choregraphe, to program the actions required to perform the task;
- NAOsim, to simulate the execution of the programmed actions with NAO considering its interaction with objects within a virtual space.

In Figure 15 the virtual space selected for the simulations is shown, which consisted of a living room. Also shown are the default positions of the objects to be manipulated by NAO as defined by the actions presented in Table 7.

Figure 15. NAOsim objects and simulation environment

The point of origin for NAO was set to be next to the bed and all actions started from the default position. The results of the Speech Interface (with the SA ASR System III) used to control the simulated NAO and perform the 20 programmed actions are presented in Table 8. The speech interface achieved a mean correct recognition rate of sentences of over 95%, which led to correct execution of associated simulated tasks.

| Test Sentence | Description of the Associated Action |
|---------------|--------------------------------------|
| 1             | Push the Small Chair to point of origin |
| 2             | Push the Small Table to point of origin |
| 3             | Push the Capsule to point of origin |
| 4             | Push the Torus to point of origin |
| 5             | Push the Cube to point of origin |
| 6             | Push the Capsule next to the Small Chair |
| 7             | Push the Torus next to the Cube |
| 8             | Push the Cube next to the Small Table |
| 9             | Push the Small Chair next to the Capsule |
| 10            | Push the Small Table next to the Cube |
| 11            | Push the Cube next to the Door |
| 12            | Push the Small Table next to the Vase |
| 13            | Push the Small Chair next to the Lamp |
| 14            | Push the Torus under NAO’s Picture |
| 15            | Push the Capsule next to the Door |
| 16            | Push the Cube + Small Chair next to the Small Table |
| 17            | Push the Torus + Small Table next to the Small Chair |
| 18            | Push the Capsule + Torus next to the Cube |
| 19            | Push all objects next to the Door |
| 20            | Push all objects next to the Vase |

Table 7. Actions associated to recognized sentences for control of a simulated NAO.

| Test User | Fails/Total | %Correct Sentences |
|-----------|-------------|--------------------|
| GJ        | 0/20        | 100                |
| MM        | 1/20        | 95                 |

Table 8. Performance of the Speech Interface for control sentences: dysarthric speech.

Although these tasks were very basic, the principle of recognizing dysarthric speech with high accuracy and controlling a simulated artificial entity has been accomplished. This can be further extended to real entities and more complex tasks.

Finally, the Assessment Module was tested. In Figure 16 and 17 the phoneme confusion matrices for GJ and MM are shown. This information was compared with perceptive tests performed by the therapists using the set of 49 words used in this module. For GJ, the therapists agreed with the significant confusions observed for the phonemes /b/, /t/, /d/, /l/, /r/ and /g/. For other phonemes such as /p/, /f/, /s/, /n/ and /h/, very few deficiencies were perceived. This agreed with the patterns shown by the phoneme confusion matrix. For speaker MM, more confusions were observed, although with less deletions and insertions. The confusions were corroborated by the therapists and although they agreed with the problems detected by the interface, they disagreed with the phonemes /b/ and /t/. Note that, although there are significant confusions, deletions and insertions, the pattern observed is not so different from the patterns of a normal speaker (see Figure 13). This could be due to the mild level of dysarthria from the speakers, where no severe abnormalities in phoneme articulation are evident.

Figure 16. Phoneme confusion patterns of speaker GJ.

Figure 17. Phoneme confusion patterns of speaker MM.
It is understood that professional assessment of dysarthria is a complex process. In addition to perceptual intelligibility tests (i.e., FDA), assessment requires visual analysis of the speech articulators (movement and strength of the lips, tongue, jaw) and analysis of oral sensitivity to tactile stimulation, breathing, etc. Thus, it is important to mention that tools based on intelligibility tests, such as the one developed for this module, only cover an aspect of the assessment of dysarthria. However, as a starting point, this module provides useful information for the task.

5. Conclusions and Future Work

In this paper our advances towards the development of a communication and assessment interface for dysarthric Mexican Spanish speakers were presented. The Speaker Adaptive (SA) approach, where the base ASR system was trained with a single-speaker phonetically-balanced normal speech corpus, seems to provide a reliable framework to accomplish high performance when resources are limited. On the other hand, the interface, by manipulation of three main variables: Gaussian components, vocabulary-LM (control of LM’s perplexity) and scale grammar factor, achieved performances of up to 95% for dysarthric speech, with dynamic adaptation being a key additional factor that contributed to this performance. This interface also showed high performance when tested with normal speech. In the HRI field, with the proposed speech interface, simulated execution of basic tasks was achieved with a mean rate over 95% for dysarthric speakers.

The results obtained give confidence about the feasibility of the interface and the levels of performance that it can achieve in real-time use. Nevertheless, more research is in progress and as future work we suggest the following:

- to evaluate the performance of the SA ASR system for larger vocabularies (near 1000 words) and dynamic adaptation;
- to analyse the effect of perplexity control and dynamic adaptation for more severe dysarthric speakers;
- to refine the classification of phonemes for the assessment module;
- to provide a measurable metric for the speaker’s level of dysarthria (i.e., provide a scale for the level of severity) from the phoneme confusion-matrix;
- to adapt the interface for its implementation in mobile devices such as smartphones;
- to increase the size of the training speech corpus of dysarthric speech;
- to improve or adapt the interface for users with additional disabilities (as in the case of GJ) and its usability;
- to integrate the interface into a more complex system to interact with a real NAO or other robotic platform for assistive tasks;
- to extend the development of the speech interface to include a module to extract semantic meaning of sentences to execute tasks (e.g., a dialog system) and improve HRI.

6. Acknowledgments

We want to thank to Dr. Mara Luisa Gutierrez (SNDIF coordinator) and therapists Rocío Bazan Pacheco (language therapy) and Diana Perez Hernandez (occupational therapy) for their support through the development of this work.

7. References

[1] Odashima, T., Onishi, M., Riken, N., Hirano, S., Mukai, T., Luo, Z. (2008) Development of the Tactile Sensor System of a Human-Interactive Robot “RI-MAN”, IEEE Transactions on Robotics, Vol. 24, No.2, pp. 505-512.
[2] PAL Robotics (2011) REEM – Humanoid Robot. Available: http://www.pal-robotics.com/robots/reem Accessed 2012 July 20.
[3] Nimbro@Home (2012) Robots: Cosero. Available: http://www.nimbro.net/Home/robots.html. Accessed 2012 July 20.
[4] Instituto Nacional de Estadística y Geografía – INEGI. Discapacidad en México (2010). Available: http://cuentame.inegi.org.mx/poblacion/discapacidadd.asp?tema=P, Accessed 2012 July 20.
[5] Tapus, A., Matarić, M. (2006) Towards Socially Assistive Robotics. Int. Journal of the Robotics Society of Japan, Vol. 24, No. 5, pp. 576-578.
[6] Lizandra, R. (2007) Dificultades en el Desarrollo del Lenguaje Oral e Intervención. National University of Madrid, Spain. Available: http://www.uam.es/personal_pdi/psicologia/agonzale/Asun/2007/DF/Artic/LizandraDificultades_lenguaje_oral.pdf. Accessed: 2012 July 20.
[7] Moriana, M. (2009) La disartria. Revista Digital Innovación y Experiencias Educativas, No. 16, pp. 1-18.
[8] Darley, F., Aronson, A., Brown, J. (1975) Motor Speech Disorders. Saunders. 304p.
[9] Ferrier, L., Shane, H., Ballard, H., Carpenter, T., Benoit, A. (1995) Dysarthric speaker’s intelligibility and speech characteristics in relation to computer speech recognition. Augmentative and Alternative Communication, Vol. 11, pp. 165-175.
[10] Manasse, N., Hux, K., Rankin-Erickson, J. (2000) Speech recognition training for enhancing written language generation by a traumatic brain injury survivor. Brain Injury, Vol. 14, pp. 1015-1034.
[11] Manasse, N., Hux, K., Rankin-Erickson, J., Lauritzen, E. (2000) Accuracy of three speech recognition
systems: Case study of dysarthric speech. Augmentative and Alternative Communication, Vol. 16, pp. 186-196.
[12] Raghavendra, P., Rosengren, E., Hunnicutt, S. (2001) An investigation of different degrees of dysarthric speech as input to speaker adaptive and speaker dependent recognition systems. Augmentative and Alternative Communication, Vol. 17, pp. 265-275.
[13] Jayaram, G., Abdelhamied, K. (1995) Experiments in dysarthric speech recognition using artificial neural networks. Journal of Rehab. Research and Development, Vol. 42, pp. 162-169.
[14] Strik, H., Sanders, E., Ruiter, M., Beijer, L. (2002) Automatic recognition of Dutch dysarthric speech: A pilot study. In: Proc. of ICSLP 2002. pp. 661-664.
[15] Matsumasa, H., Takiguchi, T., Ariki, Y., Li, I., Nakabayashi, T. (2008) Integration of metamodel and acoustic model for speech recognition. In: Proc. of Interspeech 2008, pp. 2234-2237.
[16] Hawley, M. S., Enderby, P., Green, P., Cunningham, S., Brownsell, S., Carmichael, J., Parker, M., Hatzis, A., O’Neill, P., Palmer, R. (2007) A speech-controlled environmental control system for people with severe dysarthria. Medical Engineering & Physics, Vol. 29, pp. 586-593.
[17] Livingston, N. Hamidi, F., Baliko, M., Spalteholz, L. (2010) CanSpeak: A customizable speech interface for people with dysarthric speech. In: K. Miesenberger et al. (Eds.): ICCHP 2010, Part I, LNCS 6179, Springer-Verlag Berlin Heidelberg, pp. 605-612.
[18] Miranda, C., Camal, R., Cen, J., Gonzalez, C., Gonzalez, S., Garcia, M., Narvaez, L. (2007) Un juego de gravedad con reconocimiento de voz para niños con problemas de lenguaje. In: Proc. of Workshop on Perspectives, Challenges and Opp-ortunities for HCI in Latin America, CIILHC 2007.
[19] Copalcua, M. (2009) Sistema de procesamiento de fonemas para la rehabilitación de personas con problemas de habla mediante técnicas de aprendizaje automático. MSc. Dissertation, Technological Institute of Apizaco, México.
[20] Green, P., Hatzis, A., Parker, M., Carmichael, J., Cunningham, S., O’Neill, P., Palmer, R. (2003) An Integrated Toolkit Deploying Speech Technology for Computer Based Speech Training with Application to Dysarthric Speakers. In: Proc. of Eurospeech 2003. pp. 2213-2216.
[21] Uraga, E., Pineda, L. (2002) Automatic generation of pronunciation lexicons for Spanish. In: Proc. of Int. Conf. on Computational Linguistics and Intelligent Text Processing, CICLing 2002. A. Gelbukh (Ed.), Springer-Verlag Berlin Heidelberg, pp. 300-308.
[22] Cuévara, J. (2004) Fonética de la Ciudad de México: Aportaciones desde las Tecnologías del Habla. MSc. Dissertation, National Autonomous University of México (UNAM), México.
[23] Pineda, L., Villaseñor, L., Cuétara, J., Castellanos, H., Galescu, L., Juárez, J., Lister, J., Pérez, P. (2010) The corpus DIMEX100: Transcription and Evaluation. Language Resources and Evaluation, Vol. 44, pp. 347-370.
[24] Montalto, L. (1997) Fiesta en la Montaña. Available: http://home.cc.umanitoba.ca/~fernand4/fer/tienda.html. Accessed 2012 July 20.
[25] Green, P., Hawley, M., Enderby, P., Brownsell, S., Hatzis, A., Cunningham, S., Parker, M., Carmichael, J., Palmer, R., O’Neill, P. (2003) STARDUST Speech Training and Recognition for Dysarthric Users of Assistive Technology. In: Proc. of Association for the Advancement of Assistive Technology in Europe (AAATE).
[26] Young, S., Woodland, P. (2006) The HTK Book (for HTK Version 3.4). Cambridge University, U.K.
[27] Rabiner, L. (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. In: Proc. of IEEE, Vol. 77, pp. 257-286.
[28] Jurafsky, D., Martin, J. (2009) Speech and Language Processing. Pearson: Prentice Hall.
[29] Chen, S., Beeferman, D., Rosenfeld, R. (1998) Evaluation metrics for language models. DARPA Broadcast News Transcription and Understanding Workshop.
[30] Leggetter, C., Woodland, P. (1995) Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. Computer Speech and Language, Vol. 9, pp. 171-185.
[31] Caballero, S. (2011) Automatic Recognition of Dysarthric Speech. LAP Lambert Academic Publishing. 148p.
[32] Menendez-Pidal, X., Polikoff, J.B., Peters, S.M., Leonio, J.E., Bunnell, H.T. (1996) The Nemsours Database of Dysarthric Speech. In: Proc. of the Fourth International Conference on Spoken Language Processing.
[33] Robinson, T. (1995) WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. In: Proc. of the IEEE Conf. on Acoustics, Speech and Signal Processing, pp. 81-84.
[34] Caballero, S., Cox, S. (2009) Modelling Errors in Automatic Speech Recognition for Dysarthric Speakers. EURASIP J. Adv. Signal Processing.
[35] Cannarozzi, G. (2011) String Alignment using Dynamic Programming. Available: http://www.bio recipes.com/DynProgBasic/code.html. Accessed: 2012 July 20.
[36] National Institute of Standards and Technology, NIST (2009) The History of Automatic Speech Recognition Evaluations at NIST. Available: http://www.itl.nist.gov/iad/mig/publications/ASRhitory/index.html. Accessed: 2012 July 20.