Automatic detection of articulation disorders in children with cleft lip and palate

Andreas Maier, Florian Hönig, Tobias Bocklet, and Elmar Nöth[a)]
Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen-Nürnberg, 91058 Erlangen, Germany

Florian Stelzle and Emeka Nkenke
Mund- Kiefer- und Gesichtschirurgische Klinik, Universitätsklinikum Erlangen, 91054 Erlangen, Germany

Maria Schuster
Abteilung für Phoniatrie und Pädaudiologie, Universitätsklinikum Erlangen, 91054 Erlangen, Germany

(Received 5 November 2008; revised 1 August 2009; accepted 6 August 2009)

Speech of children with cleft lip and palate (CLP) is sometimes still disordered even after adequate surgical and nonsurgical therapies. Such speech shows complex articulation disorders, which are usually assessed perceptually, consuming time and manpower. Hence, there is a need for an easy to apply and reliable automatic method. To create a reference for an automatic system, speech data of 58 children with CLP were assessed perceptually by experienced speech therapists for characteristic phonetic disorders at the phoneme level. The first part of the article aims to detect such characteristics by a semiautomatic procedure and the second to evaluate a fully automatic, thus simple, procedure. The methods are based on a combination of speech processing algorithms. The semiautomatic method achieves moderate to good agreement ($\kappa \approx 0.6$) for the detection of all phonetic disorders. On a speaker level, significant correlations between the perceptual evaluation and the automatic system of 0.89 are obtained. The fully automatic system yields a correlation on the speaker level of 0.81 to the perceptual evaluation. This correlation is in the range of the inter-rater correlation of the listeners. The automatic speech evaluation is able to detect phonetic disorders at an experts'level without any additional human postprocessing.
© *2009 Acoustical Society of America.* [DOI: 10.1121/1.3216913]

PACS number(s): 43.70.Dn, 43.72.Ar, 43.80.Qf, 43.80.Vj [DOS] Pages: 2589–2602

I. INTRODUCTION

Communication disorders are a major challenge in the 21st century because of their personal and financial impact. The cost of care as well as the decrease in employment opportunities for people with communication disorders cause a loss of $154 billion to $186 billion per year to the economy of the United States of America alone.[1] People with speech disorders do not only suffer from restricted speech but also from vocational limitations. The use of automatic speech processing techniques can contribute to reduce the associated costs. More specifically, such methods can affect speech screening and therapy as follows.

- Speech processing can serve as an easy-to-apply diagnostic tool and can also be used for speech screening. The cost of diagnosis can be reduced with such an automatic system because it can also be performed by nonprofessionals.
- Therapy strategies can be evaluated and compared against each other in clinical trials or for individual therapy.
- Speech processing can support therapy sessions in the practice as well as telemedical therapy sessions, which can be performed by the patient from his home.

In this work we focus on speech attributes related to cleft lip and palate (CLP). CLP might cause communication disorders, especially articulation disorders. CLP is the most common malformation of the head. It constitutes almost two-thirds of the major facial defects and almost 80% of all orofacial clefts.[2] Its prevalence differs in different populations. CLP appears most often in Asians with a prevalence of 1 in 400–500 newborns and least often in African Americans with 1 in 1500–2000 newborns.[3,4] Speech of children with CLP is sometimes still disordered even after surgery and might show special characteristics such as hypernasality (HN), backing, and weakening of consonants.[5]

The major feature of disordered speech in CLP is HN in vowels (perceived as characteristic "nasality") and nasalized consonants (NC). This may reduce the speech intelligibility.[6–8] Both features, HN and NC, can be summarized as nasal air emission.

The term nasality is often used in the literature for two different kinds of nasality: HN and hyponasality. While HN is caused by enhanced nasal emissions, as in CLP children, hyponasality is caused by a blockage of the nasal airway, e.g., when a patient has a cold. There are several studies on both nasality types.[9] However, most of them concern only the effects on voiced speech (vowels)[10–12] and consonant-vowel combinations.[13,14]

Figure 1 shows the effect of nasalization in the envelope spectrum[15] of vowel /a:/. In both spectra a slight nasal formant $F_1^N(f)$ exists between at frequency $f=300$ and 500 Hz.

[a)]Author to whom correspondence should be addressed. Electronic mail: andreas.maier@cs.fau.de

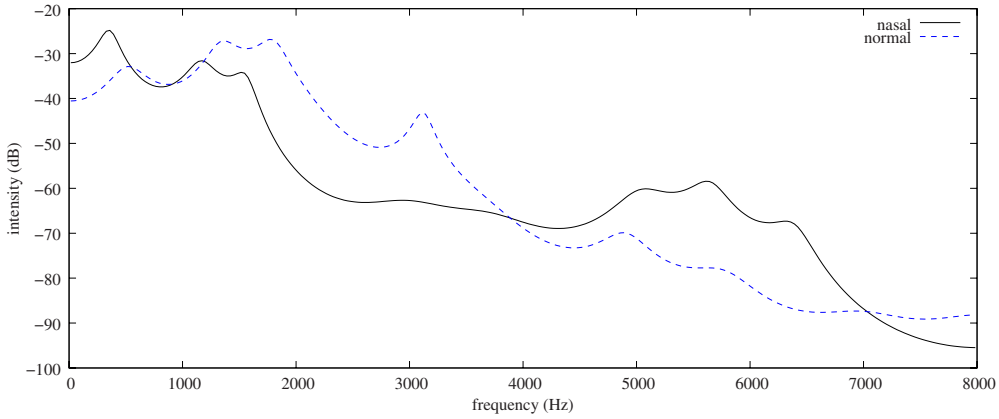

FIG. 1. (Color online) LP-model spectrum of a nasal and a non-nasal realization of the phoneme /a:/ in the phonetic context /ha:s@/ ("Hase," the German word for "hare") using 20 LPC-coefficients: The intensity of the nasal formant $F_1^N(f)$ ($f \approx 300-500$ Hz) is stronger than the intensity of the first formant $F_1(f)$ ($f \approx 1100-1300$ Hz) in the nasal realization. Note that the displayed speech is children's speech, which causes exceptionally high formant frequencies.

The maximal intensity of the first formant $F_1(f)$ is at about 1100–1300 Hz. In the nasalized /a:/, the intensity of the $F_1^N(f)$ is stronger than the $F_1(f)$, which makes the nasality audible. Actually, this effect is caused by a combination of the following effects.[16]

- The first formant bandwidth increases while the intensity decreases.
- The nasal formant $F_1^N(f)$ emerges or is increased.
- Antiresonances appear, which increase the strength of the so-called antiformants $F_k^A(f)$.

According to the literature, the main cause for audible nasality is the intensity reduction in the first format.[16,17]

Nasality in consonants, however, shows different acoustic properties depending on their mode of articulation, e.g., voiced or unvoiced. Effects in the formant structure can only be analyzed in the neighboring vowels. The effects on the consonants, however, are still audible. In fricatives, for example, the nasality is audible as a general weakening of the energy of the phoneme with additional streaming noises caused by the nasal air flow. In contrast to the non-nasal consonant the way to the nasal cavity is open. Hence, at least some of the emitted air flows through the nose and the amount of air that is emitted through the mouth is reduced. In the literature these effects are rarely described and often only the analysis of vowel-consonant clusters is performed.[18]

The speech of CLP children might also contain secondary cleft-type characteristics. These originate from compensatory articulation, which may still be present even after adequate closure of the clefting. For example, pharyngeal backing (PB) is caused by a shift in the localization of the tongue toward the palate during the articulation. Glottal articulation [also called laryngeal replacement (LR)] is an extreme backing of articulation. The resulting acoustic realization is similar to that of a glottal stop. Another typical characteristic of secondary phonetic disorders is the absence or weakening of consonants[5] [weakened plosives (WPs)].

In clinical practice, articulation disorders are mainly evaluated perceptually, and the evaluation procedures are mostly performed by a speech therapist. Previous studies have shown that experience is an important factor that influences the judgment of speech disorders. The perceptual evaluation of persons with limited experience tends to vary considerably.[19,20] For scientific purposes, usually the mean score judged by a panel of experienced speech therapists serves as a reliable evaluation of speech and is sometimes called "objective." Of course, this is very time and manpower consuming. Until now, objective measures only exist for nasal emissions[7,9] and for voice disorders in isolated vowels.[17,21] But other specific articulation disorders in CLP cannot be reliably and objectively quantified yet. In this paper, we present a new technical procedure for the objective measurement and evaluation of phonetic disorders in connected speech, and we compare the obtained results with perceptual ratings of an experienced speech therapist. We present two experiments.

- In a first experiment an automatic speech recognition (ASR) system was applied to evaluate the detection of the above mentioned articulatory features of CLP speech (HN, NC, PB, LR, and WP). The experiment is based on the transliteration of the tests that was created manually.
- A second experiment was conducted to examine whether it is possible to perform the assessment fully automatically without manual transliteration.

II. SPEECH DATA

58 children with CLP were recorded during the commonly used PLAKSS speech test (psycholinguistische analyse kindlicher sprechstörungen — psycholinguistic analysis of children's speech disorders). The acoustic speech signal was sampled at 16 kHz with a quantization of 16 bits. Informed consent had been given by the parents prior to all recordings.

For the first experiment recordings of 26 children at an age of 9.4 ± 3.3 years were used (CLP-1). Two of the children in the data set had an isolated cleft lip, 3 an isolated cleft palate, 19 unilateral CLP, and 2 bilateral CLP. The recordings were made with a head set (dnt Call 4U Comfort) and a standard PC.

The recordings were performed in the same manner as during the therapy session: The test was presented on paper-

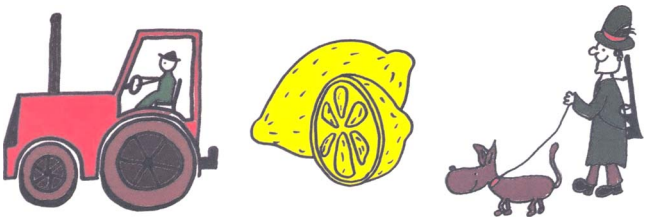

FIG. 2. (Color online) Slide 13 of the PLAKSS test: "Trecker, Zitrone, Jäger" (tractor, lemon, hunter).

board and stored in a single audio file. Therefore, the time stamps at which the therapist switched from one slide to another was not known. The data set is a subset of the data that were investigated in a previous study concerning speech intelligibility[22] and semiautomatic evaluation of speech disorders.[23]

The second group (CLP-2) was formed by 32 children at the age of 8.7 ± 1.7 years. Five of the children had a cleft lip, 7 a cleft palate, and 20 a unilateral CLP. No child in the data set had a bilateral cleft. They were recorded directly at the PC. The PC was used to display the slides and to perform the recording simultaneously. The audio data of each slide are stored in an individual audio file. Hence, the correspondence of audio data and the respective slide is clear. Furthermore, we presented the correct target words in small letters at the bottom of the screen in order to decrease the variability in the test data. The data were recorded and evaluated as described in the following paragraph using the program for the evaluation of all kinds of speech[24] (PEAKS) disorders.

The PLAKSS test[25] is a German semistandardized test for articulation disorders. It consists of 99 pictograms (97 disjoint) which are shown on 33 slides. It was designed to record also young children who are not yet able to read. The test contains all German phonemes in different positions (word initial, central, and final).

Figure 2 shows an example of the slides. It depicts the German words "Trecker, Zitrone, Jäger" to test for the phoneme /r/ in consonant-consonant clusters and at the end of a word. The words mean tractor, lemon, and hunter in English. It gives a good example: While the tractor and the lemon are quite easy to identify, the hunter often poses a problem. Many children do not recognize the rifle on the back of the hunter and call the pictogram "man with a dog." Furthermore, the word "Trecker" is rather uncommon in the southern part of Germany. Children tend to prefer variants such as "Traktor" or "Bulldog." Therefore, the vocabulary of the PLAKSS test has to be extended with common word alternatives and regional variants if their automatic detection is desired.

As the test has to be performed by a supervisor who gives instructions during the test, the voice of the supervisor is always also audible on the audio tracks.

III. SEMI- AND FULLY AUTOMATIC SEGMENTATION

In both data sets CLP-1 and CLP-2 the data were segmented using an ASR system. We use an ASR system based on hidden Markov models (HMMs). It is a word recognition (WR) system developed at the Chair of Pattern Recognition (Lehrstuhl für Mustererkennung) of the University of Erlangen-Nuremberg. In this study, the latest version[24] was used.

As the performance of speech recognition is known to be dependent on age,[26] several recognizers were trained for certain age groups. According to previous evaluations,[27] the best groups for the creation of age-dependent recognizers were found to be

- <7 years,
- 7 years,
- 8 years,
- $9+10$ years, and
- >10 years.

A maximum likelihood linear regression (MLLR) adaptation was performed on the acoustic models using the HMM output probabilities[28–30] in order to improve the recognition for each child.

The CLP-1 data set was segmented semiautomatically using the transliteration of the speech data. In the CLP-2 database this step was replaced by a fully automatic procedure using PEAKS.[24] These segmentation procedures are described in the following.

A. Semiautomatic segmentation procedure

As the whole speech data of one child were collected in a single audio file in the CLP-1 data, the complete data set had to be transliterated in order to perform segmentation. Each word was assigned a category in order to enable the distinction of target words and additional carrier words. The categories consisted of the 97 target words of the PLAKSS test plus an additional category "carrier word" for additional words that are not part of the test vocabulary. In the recordings of the 26 children, 2574 (26×99) target words were possible. However, only 2052 of the target words are present in the transliteration. This is related to the fact that the test was presented in pictograms. Hence, many children used alternatives to describe the pictogram. Sometimes, children also failed in the identification of a pictogram. As the test is rather long especially for young children with speech disorders, the therapist did not insist on the correct realization for each pictogram. She also counted alternatives as correct in order to keep the child motivated throughout the test.

In the next step the ASR system was used to segment the CLP-1 data into words and phones. All carrier words were excluded in the subsequent processing. Another 136 target words could not be used because the automatic segmentation failed, i.e., the segmented word was shorter than 100 ms. Hence, 93.3% of the appearing target words could be successfully segmented.

At the end of the semiautomatic segmentation procedure, 1916 words and 7647 phones, all from the target words, were obtained. This corresponds to 74.4% of the 2574 possible target words.

B. Fully automatic segmentation procedure

For the CLP-2 data set the semiautomatic segmentation procedure was replaced by a fully automatic one. Since the

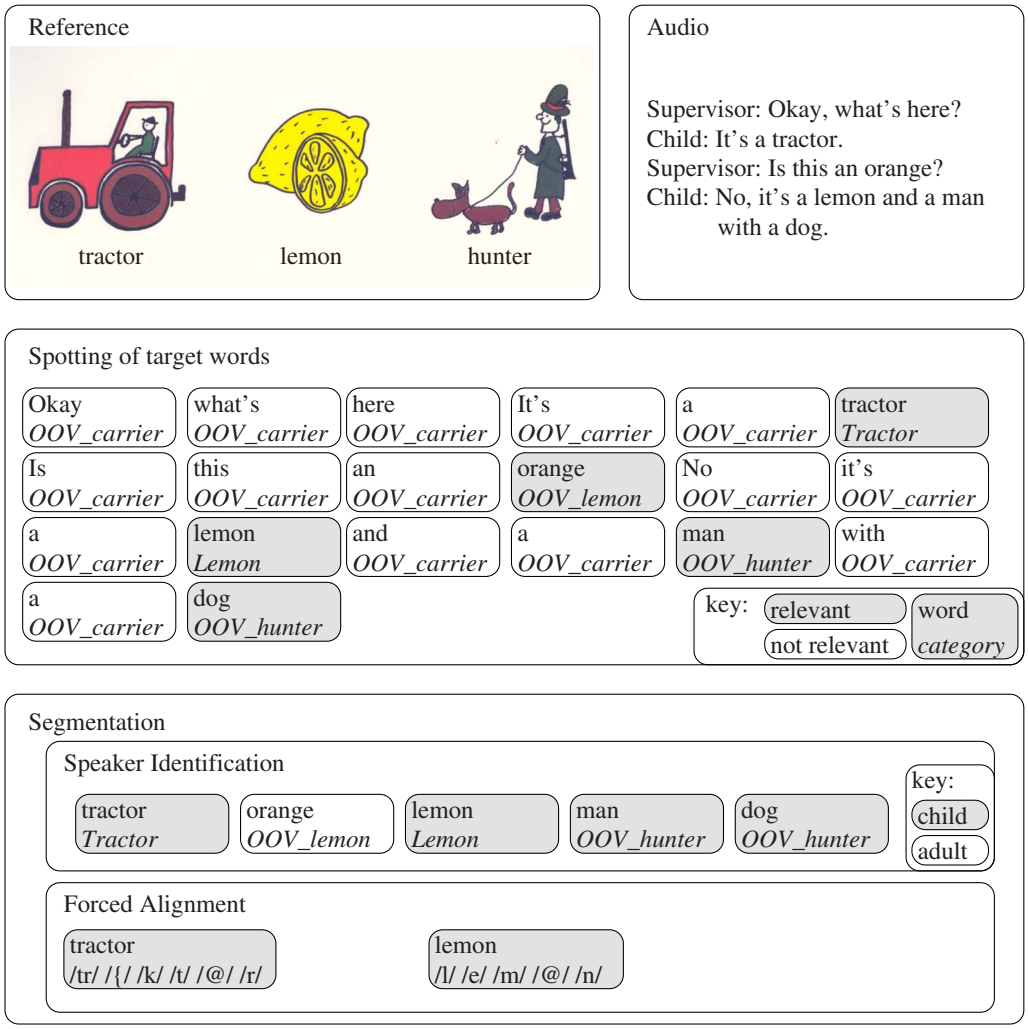

FIG. 3. (Color online) Diagram of the fully automatic word spotting and segmentation procedure: The audio data are processed by a speech recognition system. Its output is a chain of word and word category pairs. OOV words can also be detected. The category is used to identify the target words of the test. "Carrier words" are excluded from the subsequent processing. Alternatives of target words that still include the phonemic target are allowed. In the segmentation step, the speaker of each word (either the child or the therapist) is identified via energy thresholding. Finally, the successfully identified and categorized words are processed using forced alignment.

uttered word chain is not known *a priori* (cf. Sec. II), segmentation is much more difficult than in read speech, where the reference is known. First, candidates for target words have to be spotted. We do this using multiple ASR systems that are fused on word lattice level. Based on this recognition result, the target words are then extracted. Figure 3 shows a diagram of this processing, which is explained in more detail in the following.

In order to improve the segmentation, a speech recognition system with multiple trigram language models is used. The language models were created using the transliteration of the speech tests of 262 children with and without speech disorders. The categories of the language model were the 97 distinct words of the employed speech test, plus an additional category for words that appear in the "carrier sentences." In order to enable the recognition of misreadings, an out-of-vocabulary (OOV) word was added to each category.

Since the speech data were transliterated according to the acoustic realization of a child, the correspondence between the spoken words and the test words is not always clear. This is caused by the use of synonyms and pronunciation errors. However, every word of the transliteration has to be assigned to a category for the training of the language models. In order to solve this problem, an alignment was performed between the transliteration of 262 previously transliterated tests (of which the CLP-1 data are a subset) and the correct sequence of words with dynamic time warping.[31] In order to improve the alignment of pronunciation variants, the distance of substitutions was calculated according to the Levenshtein distance of the two words divided by the number of letters in the longer word. The procedure still has the problem that it is not capable of modeling variations in the sequence of the words that happen when a child names the words from right to left instead from left to right. Therefore, all found correspondences were checked manually according to their plausibility. Implausible correspondences were removed. So about 20 different alternatives of each word of the test were found.

In the transliteration of the 262 speech test sessions two tendencies could be seen: Some children use many carrier words while others use none at all. Furthermore, we built one turn-dependent and one turn-independent (using all of the

TABLE I. Articulation errors were annotated in the data by an experienced speech therapist according to the definition of Sell *et al.* (Ref. 5) in the group of 26 CLP children (CLP-1).

Speech disorder (criterion)	Description	Abbreviation
Hypernasality in vowels	Nasal air flow throughout the vowel	HN
Nasalized consonants	Air is emitted through the nose during the articulation of the consonants	NC
Pharyngealization	Tongue is shifted backward toward the pharynx during articulation	PB
Glottal articulation (laryngeal replacement)	Plosives are sucked to the larynx	LR
Weakened pressure consonants	Articulatory tension is diminished	WP

transliteration) model each. The turn-dependent models used only words that actually appeared in the transliteration of the processed turn to decrease the variability in the recognition. The turn-independent models were trained using all of the transliterations that were available and therefore allowed more variability. The segmentation is then performed using four language models for each turn: Two (one turn-dependent and one turn-independent) were trained on sentences with two or more carrier words per slide, and another two models with two or fewer carrier words. In preliminary experiments, trigram language models proved to yield the best recognition rates (RRs) in all four cases compared to language models with larger or smaller context.[27]

To estimate the probability of the OOV words, each word that occurred fewer than three times was used to train the OOV language model probabilities. The probabilities of the OOV words in the language model were estimated using the VOCSIM algorithm.[32] The acoustic realization of the OOV words is flat, i.e., it is assumed to be any sequence of the phonemes of the speech recognizer.

The recognition was performed for each turn using four different language models as described above. In order to obtain a single word chain, the four best word lattices plus the reference lattice, i.e., the actual object names, were merged using the recognizer output voting error reduction.[33–35]

In this manner, an improved recognized word chain is obtained. Preliminary experiments[27] were performed using the database with the 262 children. The data were split into training and a test set. All of the 26 children of the CLP-1 data were part of the test set. An increase in the word accuracy (WA) (cf. Sec. V A) of normal children speech from 64.7% to 74.5% was found. In the CLP speech data, this improvement was even more evident. The WA of −11.0% of the baseline system without any adaptation was pushed to 42.6%.

From the 3128 (32×99) target words that appear in the CLP-2 data, 2981 could be successfully extracted from the data. This corresponds to 94.0% of all target words. This percentage is much higher than in the semiautomatic case because the correct target names were shown below the pictogram. If we take a look at the successful segmentation rate of the semiautomatic system only, both are comparable (93.3% in the semiautomatic case).

IV. PERCEPTUAL EVALUATION OF THE SPEECH DATA

A speech therapist with many years of specialized experience thoroughly evaluated the CLP-1 data set. She differentiated all criteria, as listed in Table I. The therapist evaluated all target words that appeared in the transliteration by marking each affected phone.

Two other speech therapists examined the feature "nasal emission" as the most frequent error of the CLP-1 subset with implicit differentiation of nasalized vowels and NCs on phoneme level. They marked each phone either as "nasal" or "non-nasal." Due to the automatic segmentation procedure, only the target words of the PLAKSS test that could be segmented automatically were evaluated.

V. AUTOMATIC SPEECH DISORDER EVALUATION SYSTEM

The automatic evaluation system is divided into preprocessing, feature extraction, classification, and results and concludes with a decision for a specific class.[36] A scheme is shown in Fig. 4. The entire procedure is performed on the frame, phoneme, word, and speaker levels. On each of these levels different state-of-the-art features are computed.

On the frame level, mel frequency cepstrum coefficients (MFCCs) hold relevant information for the articulation. As features on the phoneme level, we extract teager-energy-profile (TEP) features as they have been shown to be relevant for the detection of nasality in vowels.[18] Furthermore, we compute pronunciation features on the phoneme level (*PronFexP*) as they were successfully applied to pronunciation scoring of non-native speech.[37] On the word level, the pronunciation features of Hacker *et al.*[38] have also been shown to be effective for the assessment of the pronunciation of second language learners. Also prosodic features (*ProsFeat*) may hold relevant information on the speech characteristics.[39] Hence, they were also included in the assessment procedure on the word level. On the speaker level, i.e., using all audio data of the speaker without segmentation, we included Sammon features[40] and the recognition accuracy of an ASR system[22] as both have been shown to be correlated to the speech intelligibility. Table II reports a summary of these features.

In our classification system we apply the concept of "late fusion,"[41] i.e., we train a classifier for each level. Com-

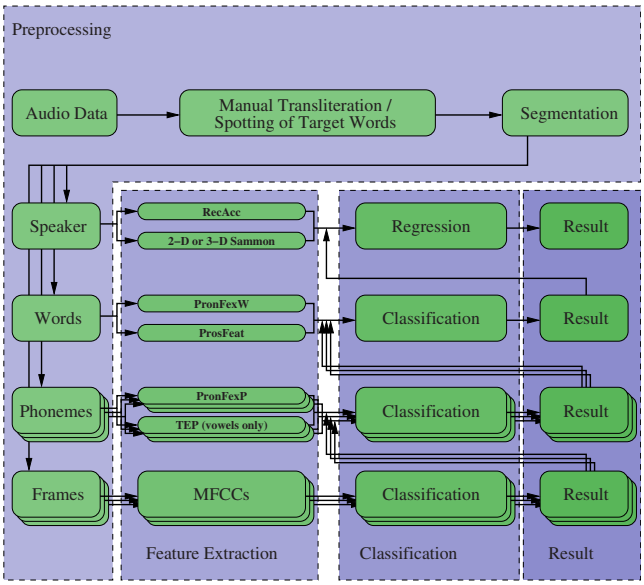

FIG. 4. (Color online) Experimental setup of the classification system: Right after the recording the preprocessing is performed. The data are transliterated manually or spotted automatically and segmented using forced alignment. Next, the feature extraction takes place on multiple levels. The features are supplied to a classifier that performs evaluation afterward. The output of each classifier is used as feature on the respective next evaluation level. On the frame and the phoneme levels, an independent classifier is trained for each phoneme. This is denoted as parallel arrows in the figure.

pared to other classification tasks, we do not have the option of "early fusion," i.e., concatenation of feature vectors, as this procedure would end in feature vectors of different length in vowels and consonants as some features can only be computed in vowels. Hence, we train different classifiers on the frame and phoneme levels for each phoneme. The output of each classifier is then used as feature on the respective next level. This is represented in Fig. 4 as multiple parallel arrows.

From the corresponding result of the lower level, features are computed and supplied to the higher level. These features include the mean, the maximum, the minimum, the standard deviation, the sum, and the product of the output probabilities. Furthermore, the absolute and relative frequencies of the decision for each class are regarded as features. Note that no information about the actual class membership is included in this process.

It is possible to compute evaluation results on all levels. We report these numbers to give an impression of the importance of the different feature groups on the respective levels. The main focus of this article, however, is the evaluation result on the speaker level.

For details on MFCCs, see, for example, Ref. 42. All other features are reported in detail in Secs. V A–V F. The section on the automatic speech disorder evaluation system is concluded by a description of the classifiers.

A. Recognition accuracy features

Good correlations between the intelligibility and the recognition accuracy have already been reported.[43] In our procedure we use the WR system described in Sec. III for processing the speech of the children. Then, the recognized word chain is compared to the reference, i.e., the target words, in order to determine the recognition accuracy.

In contrast to the segmentation procedure, we used a unigram language model to weigh the outcome of each word model. It was trained with the target words of the tests. Thus, the frequency of occurrence of each word in the used text was known to the recognizer. This enhances recognition results by including linguistic information. However, for our purpose it was also necessary to put more weight on the recognition of acoustic features. A comparison between unigram and zerogram language models was previously conducted.[44] It was shown that intelligibility can be predicted using WR accuracies computed by either zero- or unigram language models. The unigram, however, is computationally more efficient because it can be used to reduce the search space. The use of higher n-gram models was not beneficial in terms of correlation.[45]

For the evaluation of the recognized word chain, two measures are commonly used: the WR rate and the WA.

$$\mathrm{WR} = \frac{C}{R} \times 100\,\% \,.$$

TABLE II. Overview on the feature sets which are extracted on four different evaluation levels.

Label	Level	No.	Description	Reference
RecAcc	Speaker	2	Accuracy of the speech recognition (word correctness and accuracy)	22
2D Sammon Coordinates	Speaker	2	Coordinates on a 2D Sammon map	40
3D Sammon Coordinates	Speaker	3	Coordinates on a 3D Sammon map	40
ProsFeat	Word	37	Features based on the energy, the F_0, pauses, and duration to model the prosody of the speaker	39
PronFexW	Word	7	Pronunciation features (PronFex) to score the correctness of the current word	38
PronFexP	Phoneme	6	Features to score the correctness of the Pronunciation (PronFex) of the current phoneme	37
TEP	Phoneme	1	Teager energy profile to detect nasality in vowels	18
MFCCs	Frame	24	Mel frequency cepstrum coefficients	42

WR is computed as the percentage of correctly recognized words C and the number of reference words R. In addition,

$$\mathrm{WA} = \frac{C - I}{R} \times 100\%$$

weighs the number of wrongly inserted words I in this percentage. The WA punishes the insertion of additional words compared to the reference chain. Hence, it is known to be sensitive to carrier words.[46] The upper limit of both measures is 100%. The lower bound of the WR is 0% while the WA does not have a lower bound. It becomes negative, as soon as the recognizer inserts more wrong additional words than it actually recognizes correctly. This feature is used to support the assessment on the speaker level.

B. Sammon mapping

The speech data of each child are used to create speaker-dependent acoustic models. The adapted model coefficients are computed using MLLR adaptation of the speaker-independent model (cf. Sec. III). These adapted coefficients are then interpreted as a representation of the speaker with a fixed number of parameters, i.e., dimensions.

The Sammon transformation (ST) is a nonlinear method for mapping high dimensional data to a plane or a three dimensional (3D) space.[47] The ST uses the distances between the high dimensional data to find a lower dimensional representation—called map in this article. The ST preserves the topology of the original data, i.e., keeps the distance ratios between the low dimensional representations—called star here—as close as possible to the original distances. By doing so, the ST is cluster preserving. To ensure this, the function e_S is used as a measurement of the error of the resulting map [two dimensional (2D) case]:

$$e_S = s \sum_{p=1}^{N-1} \sum_{q=p+1}^{N} \frac{(\delta_{pq} - \theta_{pq})^2}{\delta_{pq}}, \tag{1}$$

with

$$\theta_{pq} = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}. \tag{2}$$

δ_{pq} is the high dimensional distance between the high dimensional features p and q stored in a distance matrix $\boldsymbol{D}$, θ_{pq} is the Euclidean distance between the corresponding stars p and q in the map, and N is the total number of stars. For the computation of the high dimensional distance between two speech recognizers we use the Mahalanobis distance.[48,49] s is a scaling factor derived from the high dimensional distances:

$$s = \frac{1}{\sum_{p=1}^{N-1} \sum_{q=p+1}^{N} \delta_{pq}}. \tag{3}$$

The transformation is started with randomly initialized positions for the stars. Then the position of each star is optimized, using a conjugate gradient descent library.[50] This method is referred to as comprehensive space map of objective signal by Nagino and Shozakai.[51] A further advantage of the Sammon transform is that the derived coordinates can also yield further information on the intelligibility of the speaker.[52] In our experiments, e_S was 9% of the high dimensional distances. Hence, these features can be used on the speaker level for the assessment.

C. Prosodic features

The prosody module takes the output of our WR module in addition to the speech signal as input. In this case the time-alignment with the Viterbi algorithm of the recognizer and the information about the underlying phoneme classes (e.g., *long vowel*) are used by the prosody module.[53]

First, the prosody module extracts so-called base features from the speech signal. These are the energy, the fundamental frequency (F_0) after Bagshaw *et al.*,[54] and the voiced and unvoiced segments of the signal. In a second step, the actual prosodic features are computed in order to model the prosodic properties of the speech signal. For each word point, we extract 21 prosodic features. These features model F_0, energy, and duration. In addition, 16 global prosodic features for the whole utterance, i.e., slide, are calculated. They cover each of mean and standard deviation for jitter and shimmer,[55,56] the number, length, and maximum length each for voiced and unvoiced sections, the ratio of the numbers of voiced and unvoiced sections, the ratio of length of voiced sections to the length of the signal, and the same for unvoiced sections. The last global feature is the variance of the fundamental frequency F_0. Batliner *et al.*[53] presented a more detailed description of these features.

D. Pronunciation features

Articulation disorders result in pronunciation errors. Some of these errors concern confusion of phonemes. This is similar to the misarticulations that occur with speakers of a foreign language. Therefore, the investigation of methods that were developed for the scoring of non-native speech seems beneficial. Pronunciation features[57] are used to measure the progress in learning a foreign language.[58] In this work, we study these features' applicability to the detection of pathologic speech. More precisely, we only analyze a subset of these features that is based on phoneme confusion probabilities on a word level. To calculate these phoneme confusion features, we compare the result of the forced alignment with the Viterbi algorithm of every word to the result of a phoneme recognizer. The phoneme recognizer uses semi-continuous HMMs and a 4-g language model. It is based on MFCCs calculated every 10 ms with a frame size of 16 ms (cf. Sec. III). From the reference word and the recognized phoneme chain a confusion matrix $\boldsymbol{C}$ is built. It contains for every pair of phonemes a, b the probability that a was detected by the recognizer when there should be b according to the forced alignment

$$c_{ab} = P(a|b), \tag{4}$$

where c_{ab} is the corresponding entry of matrix $\boldsymbol{C}$. From the training set, we calculate two confusion matrices: one for the pathologic speech data and one for the normal data. From these framewise results, we calculate the following features for the phoneme level.[59]

- Goodness of pronunciation:[60] Score computed from the framewise score of a forced alignment and the likelihood obtained by a phoneme recognizer that was trained with normal speech. In non-native speech, the likelihood is known to drop in mispronounced phonemes. We expect the same for pathologic speech.
- Duration score: Likelihood of the observed phoneme duration given the duration distribution observed in normal speakers.
- Acoustic score: Posterior probability of the speech recognizer for the current phoneme.
- Actual duration: Observed duration.
- Expected duration: Mean value of the duration distribution observed in normal speakers.
- Confidence score Q:

$$Q = \frac{P_{\text{pathologic}}(a|b)}{P_{\text{normal}}(a|b)}. \tag{5}$$

For the word level, the following features are extracted:[37]

- PC1: Mean of Q,
- PC2: Maximum of Q,
- PC3: Minimum of Q,
- PC4: Variance of Q,
- PC5: Median of Q,
- A1: Phoneme correctness, and
- A2: Confidence score of the recognized word computed by the speech recognizer (cf. Sec. III).

E. Teager energy profiles

The teager energy operator (TEO) is a heuristic approach of pronunciation feature extraction. The teager operator[61] has been applied to detect nasality in sustained vowels and consonant-vowel-consonant combinations.[18] The TEO is defined as

$$\psi[f(n)] = [f(n)]^2 - f(n+1)f(n-1), \tag{6}$$

where $f(n)$ denotes the time-domain audio signal. The TEO's output is called the TEP.

The TEP can be used to detect hypernasal speech because it is sensitive to multicomponent signals.[13,18] When normal speech is low-pass-filtered in a way that the maximum frequency f_{lowpass} is somewhere between the first and the second formant, the resulting signal mainly consists of the first formant. However, the same procedure with hypernasal speech results in a multicomponent signal due to the antiformant.

In order to get a reference signal that contains only the first formant a second signal is computed with a band-pass filter around the first formant. The bandpass filter covers the frequency range ± 100 Hz around the first formant. For both signals the TEP is computed and compared. We measure that difference with the correlation coefficient between both TEPs. The values with the best results for f_{lowpass} were determined experimentally.[62]

F. Classification

For the classification various algorithms as provided in the Waikato environment for knowledge analysis[63] were employed. The following classifiers were used.

- *OneR*. The classifier divides the numeric features—often called attributes in machine learning—into intervals that contain only observations—also called instances—of one class. In order to prevent overfitting, mixed intervals are also allowed. However, each interval must hold at least a given number of instances in the training data. Then a decision rule for classification is created for each attribute. At the end of the training procedure, the attribute is selected for the classification that has the highest accuracy on the training set.[64]
- *DecisionStump*. DecisionStumps are commonly used in ensemble training techniques such as boosting. The classifier selects one attribute and a threshold or decision value to perform the classification. Selection is performed with correlation in the numeric case and entropy in the nominal case. Then, the selection value with the highest classification rate on the training set is determined.
- *LDA-Classifier*. The linear discriminant analysis (LDA)-Classifier is also called "ClassificationViaRegression."[65] It basically determines a LDA feature transformation matrix and reduces the dimension to 1. Then, a simple threshold can be chosen to perform the classification. Again, the threshold is determined on the training set according to the best classification rate.
- *NaiveBayes*. The naïve Bayes classifier is trained according to the Bayes' decision rule.[36] A unimodal Gaussian mixture is often chosen as a probability density function.[66] This classifier is equivalent to a Gaussian mixture model (GMM) classifier with just one Gaussian distribution with equal prior probabilities.
- *J48*. The J48 is an implementation of a C4.5 decision tree.[67] In order to build a C4.5 decision tree, all instances in the data set are used to create a set of rules. Later on, the rules are pruned in order to reduce their number. Subsequently, a tree is generated that holds one simple decision rule concerning only one attribute, i.e., a DecisionStump in every node. At the leaves of the tree a class label is assigned. Classification is then performed starting from the tree root and following a path according to rules in the node. At the end of the classification, a leaf is reached that assigns the class to the observation.
- *PART*. In order to modify the rules for a decision tree, two dominant approaches exist. The first one is eliminating rules like the J48 tree does. The second one extends rules by replacing one or multiple rules by a better more refined rule. PART generates partial trees using both approaches and merges them later on. This method is much faster in training compared to J48 while having a similar or even better recognition accuracy.[68]
- *RandomForest*. This kind of classifier is composed of multiple trees that are created randomly. For each tree a random subset of the training data is chosen. Then, a random subset of attributes is selected to be used in the tree. At each node, features are picked at random to determine the

rule of the actual node. The rule that creates the best split for the current subset is computed. Such a random tree may not be pruned. A set of a random number of trees is then fused into a random forest.[69]

- *SVM*. A support vector machine classifier[70] tries to find a surface that separates two classes from each other. Therefore, it is not necessary to remember all observations in the training set. Only a small number of observations is really important for the classification task. More specifically, only those feature vectors that are close to the class boundary are important for the decision. Although only two-class problems are considered in our work, the SVMs can easily be adapted to multiple classes by training an individual classifier for each class. In this manner a set of two-class problems "one against all others" is created.
- *AdaBoost*. Boosting[71] is a common procedure for enhancing simple classifiers. The idea of boosting is to join many weak classifiers to one single strong classifier. This is achieved by training in several iterations. In each iteration, the data are reweighted. Previously wrongly classified instances get a higher weight while correctly classified ones get a reduced weight adapting the classifiers to the misclassified instances.

We tested each of the classifiers at every level. The use of different classifiers on different levels was also allowed, but it was not permitted on the same level, e.g., different classifiers for different phonemes. The prior distribution of the classes, e.g., the probability of a word to be marked as "hypernasal," was not changed for the classification task since we wanted to keep the experiments as realistic as possible.

VI. EXPERIMENTAL RESULTS

In the following we present the results obtained by perceptual, semiautomatic, and automatic processing. The evaluation units are frames, phonemes, and words on the respective level. All evaluation measures are computed from the confusion matrix:

$$\begin{bmatrix} TP_a & FN_b \\ FN_a & TP_b \end{bmatrix}. \tag{7}$$

TP_a is the number of true positive classifications or the observable agreement for class Ω_a, i.e., that is unit as pathologic. FN_a is the number of false negatives, i.e., that the unit is wrongly assigned to the opposite class. This can also be referred to as the observable disagreement of class Ω_a. TP_b and FN_b are defined analogously.

The mostly used measure in classification tasks is the RR defined as

$$\mathrm{RR} = \frac{TP_a + TP_b}{N} \times 100\,\%\,, \tag{8}$$

where $N = TP_a + FN_a + TP_b + FN_b$. Furthermore, we introduce the classwise averaged RR (CL)

TABLE III. Results of the perceptual evaluation of the CLP-1 database.

	No. of affected		
Speech disorder	Phones	Words	Children
Hypernasality in vowels (HN)	49	49	4
Nasalized consonants (NCs)	329	329	15
Pharyngealization (PB)	34	33	7
Glottal articulation (LR)	32	31	4
Weakened pressure consonants (WPs)	105	105	14
Total number of units	7647	1916	26

$$\mathrm{CL} = \frac{1}{2} \times \left(\frac{TP_a}{TP_a + FN_a} + \frac{TP_b}{TP_b + FN_b} \right). \tag{9}$$

The CL is also often referred to as the unweighted average recall. The recall is defined as the number of true positives divided by the number of true positives and false negatives and is, therefore, equal to the definition of the sensitivity. Furthermore, we report the multirater-κ after Davies and Fleiss.[72]

The frame, phoneme, and word level results, however, are only intermediate results. The main focus of this article is the speaker level assessment. For each speaker we compute the percentage of pathologic words. Furthermore, we compute the percentage of detected words. Then, we measure the agreement using Pearson's correlation coefficient.[73]

A. Results of the perceptual evaluation

Table III reports the number of phones, words, and children that were affected by each of the five disorders according to the speech therapist's evaluation in the CLP-1 data set. The number of words is almost the same as the number of misarticulated phonemes since a single articulation error within a word meant that the whole word was counted as disordered. Only two words in the data set contained the same type of articulation error twice, i.e., 34 phonemes in 33 words with PB and 32 phonemes in 31 words with LR were annotated (cf. Table III). The last column shows the number of children who were affected by different disorders. While HN, PB, and LR appear in only few children, WP and NC appear in more than half of the children.

Table IV shows the agreement of the both raters of the CLP-2 data set. In the perceptual evaluation of the CLP-2 subset, the agreement of both raters was moderate. Only 127 words were marked as nasal emission by both raters. 2499 of the 2981 words were not marked as nasal emission. This

TABLE IV. Confusion matrix of the ratings by the two speech therapists for the criterion "nasal emission" on the CLP-2 database on the word level: Both raters marked 127 words as "nasal" and 2499 as "non-nasal." For 355 words the raters disagreed (κ=0.352).

Nasal emission	Nasal (rater 1)	Non-nasal (rater 1)
Nasal (rater 2)	127	203
Non-nasal (rater 2)	152	2499

TABLE V. Overview on the results of the pronunciation assessment on the frame, phoneme, word, and speaker levels for the CLP-1 data: All reported correlations (r) on the speaker level are significant at $p<0.01$.

	Semiautomatic evaluation									
	Frame			Phoneme			Word			
Criterion	CL (%)	RR (%)	κ	CL (%)	RR (%)	κ	CL (%)	RR (%)	κ	Speaker r
HN	56.8	99.0	0.564	62.9	99.0	0.627	60.6	96.9	0.596	0.89
NC	62.0	94.2	0.606	68.5	95.6	0.671	63.6	82.5	0.576	0.85
LR	59.8	99.6	0.597	69.5	99.6	0.694	63.8	98.2	0.632	0.81
PB	66.0	99.1	0.659	76.9	99.6	0.768	67.9	98.2	0.673	0.70
WP	71.1	97.8	0.708	71.1	97.8	0.707	75.8	97.8	0.745	0.82

corresponds to a true positive rate of the human rater 1 of 45.5% at a false positive rate of 7.5% taking rater 2 as the reference. Rater 2 had a true positive rate of 61.5% with a false negative rate of 5.7% taking rater 1 as the reference. κ values were 0.342 on the frame level, 0.313 on the phoneme level, and 0.352 on the word level.

In order to compare the automatic system with the perceptual evaluation, we computed both measures for each of the human raters (cf. Table IV). RR is the same for both raters, i.e., the percentage of observations where both raters agree:

$$\mathrm{RR} = \frac{127 + 2499}{2981} = 88.1\ \% .$$

CL is different for each rater. For rater 1, rater 2 is the reference:

$$\mathrm{CL(rater\ 1)} = \left(\frac{127}{127 + 152} + \frac{2499}{2499 + 203} \right) \Big/ 2 = 69.0\ \% ,$$

and for rater 2 rater 1 becomes the reference:

$$\mathrm{CL(rater\ 2)} = \left(\frac{127}{127 + 203} + \frac{2499}{2499 + 152} \right) \Big/ 2 = 66.4\ \% .$$

Correlation on the speaker level showed good consistency. When the percentages of marked words per speaker of both raters were compared a correlation of 0.80 was obtained.

B. Results of the automatic evaluation

All automatic evaluation experiments on the frame, phoneme, and word levels were conducted as leave-one-speaker-out evaluation, i.e., the training of the classifiers was performed with all but one speaker who was then employed as test speaker. This process was performed for all speakers.

To obtain a reference to build the automatic system, the label nasal emission is assigned if both raters agreed on their decision on the label in the CLP-2 data. Everything else was considered to be non-nasal. As reference on the speaker level, the percentage of marked words was chosen.

As reported in Table V very high values are reached for RR for the CLP-1 data set. This, however, is related to the unbalanced test sets: Most samples in the test set are not pathologic. Hence, classification of all samples to the class "normal" already yields high RRs. In order to optimize the CL rate, the training samples were weighted to form a balanced training set. The CL shows that the accuracy is moderate in most cases for these two class problems. The κ values are lower than in the semiautomatic case ($\kappa \approx 0.45$). This is related to the moderate agreement of the two raters (cf. Table IV; $\kappa \approx 0.35$), which is used in the multirater-κ computation. If we regard only the reference which was actually shown to the classifier in the training, κ lies in the same range as in the semiautomatic case ($\kappa \approx 0.6$).

On the speaker level, the features RecAcc and 2D or 3D Sammon coordinates (cf. Table II) were added to the evaluation procedure. Significance tests revealed that all reported correlations are highly significant with $p<0.01$. Except for WP, the result of the semiautomatic system achieves correlations above 0.81 for the phonetic disorders.

On the CLP-2 data, only the criterion nasal emission was evaluated (cf. Table VI). Again, high RRs were found in all classification experiments. As in the CLP-1 data, this is related to the bias in the distribution of the classes. The CL on the frame level is lower than the CLs for HN and NC in the CLP-1 data. On the phoneme level, this difference is already compensated. The CL of 64.8% is in between the recognition results of the HN and NC criteria. The same result can be observed on the word level. On the speaker level, a high correlation to the perceptual evaluation of the

TABLE VI. Overview on the results of the fully automatic pronunciation assessment on the frame, phoneme, and word levels for the CLP-2 data. The reported κ values are computed using the multirater-κ. The κ values in parentheses are computed using just the reference and the outcome of the automatic system. The correlation on the speaker level was significant with $r=0.81$ and $p<0.01$.

	Fully automatic evaluation		
Measure	Frame	Phoneme	Word
CL	52.6%	64.8%	62.1%
RR	98.8%	98.8%	94.0%
κ	0.431 (0.521)	0.478 (0.645)	0.482 (0.605)

TABLE VII. Detailed results for the different features on the frame, phoneme, and word levels for the CLP-1 and the CLP-2 data. If more than one rater was available (CLP-2 data only), κ values in parentheses report the agreement between the automatic system and the reference only.

Disorder	Feature	Level	CL (%)	RR (%)	κ
HN	MFCCs	**Frame**	**56.8**	**99.0**	**0.564**
HN	MFCCs	Phoneme	56.9	99.0	0.566
HN	TEP	Phoneme	59.2	97.7	0.589
HN	MFCCs+TEP	**Phoneme**	**62.9**	**99.0**	**0.627**
HN	MFCCs+TEP+PronFexP	Phoneme	60.6	98.7	0.603
HN	MFCCs	Word	52.3	96.9	0.511
HN	MFCCs+TEP	Word	57.7	95.8	0.566
HN	MFCCs+TEP+PronFex	**Word**	**60.6**	**96.9**	**0.596**
HN	MFCCs+TEP+PronFex+ProsFeat	Word	56.8	62.0	0.557
NC	MFCCs	**Frame**	**62.0**	**94.2**	**0.606**
NC	MFCCs	Phoneme	66.7	94.6	0.653
NC	PronFexP	Phoneme	67.5	91.5	0.661
NC	MFCCs+PronFexP	**Phoneme**	**68.5**	**95.6**	**0.671**
NC	MFCCs	**Word**	**63.6**	**82.5**	**0.576**
NC	MFCCs+PronFex+ProsFeat	Word	58.4	62.9	0.515
LR	MFCCs	**Frame**	**59.8**	**99.6**	**0.597**
LR	MFCCs	**Phoneme**	**69.5**	**99.6**	**0.694**
LR	MFCCs+PronFexP	Phoneme	65.3	92.6	0.652
LR	MFCCs	**Word**	**63.8**	**98.2**	**0.632**
LR	MFCCs+PronFex	Word	60.0	81.1	0.594
LR	MFCCs+PronFex+ProsFeat	Word	57.7	72.6	0.570
PB	MFCCs	**Frame**	**66.0**	**99.1**	**0.659**
PB	MFCCs	Phoneme	66.7	99.6	0.666
PB	MFCCs+PronFexP	**Phoneme**	**76.9**	**99.6**	**0.768**
PB	MFCCs	Word	59.8	98.2	0.591
PB	MFCCs+PronFex	**Word**	**67.9**	**98.2**	**0.673**
WP	MFCCs	**Frame**	**71.1**	**97.8**	**0.708**
WP	MFCCs	**Phoneme**	**71.1**	**97.8**	**0.707**
WP	MFCCs+PronFexP	Phoneme	71.0	88.5	0.706
WP	MFCCs	Word	66.1	97.8	0.642
WP	MFCCs+PronFex	Word	67.7	70.7	0.659
WP	MFCCs+PronFex+ProsFeat	**Word**	**75.8**	**97.8**	**0.745**
Nasalization	MFCCs	**Frame**	**52.6**	**98.8**	**0.431** (**0.521**)
Nasalization	MFCCs	Phoneme	62.4	98.7	0.466 (0.620)
Nasalization	MFCCs+TEP	Phoneme	62.0	98.7	0.464 (0.616)
Nasalization	MFCCs+TEP+PronFexP	**Phoneme**	**64.8**	**98.8**	**0.478** (**0.645**)
Nasalization	MFCCs	**Word**	**62.1**	**94.0**	**0.482** (**0.605**)
Nasalization	MFCCs+TEP	Word	60.2	81.8	0.472 (0.585)
Nasalization	MFCCs+TEP+Pronfex	Word	59.7	68.6	0.469 (0.580)

human raters of 0.81 is achieved. This is in the same range as the inter-rater correlation. No significant difference in the regression between nasality in vowels and the nasality in consonants was found on the speaker level ($p > 0.05$).

Table VII reports a detailed overview on the classification performance of different combinations of features. The best combinations are printed in boldface.

VII. DISCUSSION

As shown on the CLP-1 data, the system detects speech disorders on the speaker level as well as an expert. The correlations between the automatic system and the human expert for the different articulation disorders were mostly in the same range. Except for WP all correlation coefficients do not differ significantly from the best correlation of 0.89 ($p > 0.05$).

The speaker level evaluation of a fully automatic system performs comparably to two experienced listeners. The proposed system was tested for nasal emissions on the CLP-2 data. We decided for nasal emissions since they are the most characteristic and frequently occurring feature of speech of children with CLP. For our classification system, the differentiation of HN in vowels and HN in consonants does not play a significant role on the speaker level. As we train dif-

ferent classifiers for each phoneme this difference is compensated by the structure of our evaluation system on the higher evaluation levels.

For both experiments, the databases were suitable for this task since both contained a sufficient amount of normal and disordered speech data. The distribution of the classes normal and "disordered" in the test data was not adjusted, so as to create an evaluation task as realistic as possible.

Although the agreement between the human raters on the frame, phoneme, and word levels was moderate, we decided to use all data to train and test the classifiers. Selection of clear prototypical cases could, of course, improve the classification performance, as shown by Seppi *et al.*[74] However, as soon as the classifier is presented less prototypical test data, the classification performance drops significantly. Since we want to create a system that is employed in clinical routine use, we also need nonprototypical data.

In the semiautomatic system 93.3% of the target words that actually appeared in the audio data were usable for the subsequent processing. The fully automatic preprocessing procedure was able to replace this step completely. With the correct target words shown on the screen, 94.0% of them could be extracted successfully.

The system employs many state-of-the-art features and algorithms that are commonly used in pronunciation scoring of second language learners. It was shown that they also work for the evaluation of disordered speech.

Surprisingly, MFCCs alone yield high RR. We relate this to the fact that MFCCs model well human perception of speech in general. Hence, the effect of articulation disorders can also be seen in the MFCCs.

The features for transferring the classification output from one level to the next higher level are very useful. From the frame to phoneme levels, the recognition virtually always increased. On the word level, the phoneme level features also contributed to the recognition.

Combination of multiple features is beneficial on all evaluation levels, especially the pronunciation features in all articulation disorders and the TEP in the disorders concerning nasal emissions. Hence, the pronunciation features can not only model the pronunciation errors by non-natives but also articulation disorders in children. The TEP, which was previously only used in vowels and consonant-vowel combinations, showed to be applicable to connected speech as well. On the speaker level, RecAcc and Sammon coordinates increased the correlation to the perceptive evaluation. Prosodic features performed weakly in general. In most cases they did not contribute to any improvement. We relate this to the fact that the PLAKSS test is based on individual words and therefore induces only little prosody.

The employed classification toolbox provides state-of-the-art classifiers and methods for their combination. In general the tree-based classifiers, the SVMs, but also the DecisionStumps and NaïveBayes Classifiers combined with AdaBoost yielded the best performance.

On the frame and phoneme levels, CLs of up to 71.1% were reached on the CLP-1 data. On the word level the best CL was 75.8%. This is comparable to other studies concerning pronunciation scoring.[38,57,75] However, we consider these rates only as intermediate results indicative of the capabilities of the classification. Although there were errors, the classification errors are systematic. In contrast to commonly used perceptual evaluation by human listeners, results are not biased by individual experience. An automatic system therewith could provide different cleft centers with a standardized detection method for speech disorders. The classification on the word level with 75.8% CL is sufficient for a good quantification of all five disorders on the speaker level, as can be seen in the high and significant correlations (0.70–0.89). The classification system shows errors but they are consistent, i.e., the number of additional instances that are classified as disordered is similar in all speakers. The percentage of disordered events can be predicted reliably by regression.

The lowest correlation was found to be 0.70 for PB while the best correlation was 0.89 for HN in the CLP-1 data. All correlations were highly significant with $p<0.01$. In previous studies we found inter-rater correlations in the same range between human experts for the same evaluation tasks.[20]

On the CLP-2 data, CLs and RRs for experiments on the frame, phoneme, and word levels were comparable to the semiautomatic case. κ, however, was reduced. This is caused by the moderate inter-rater agreement between the two human raters ($\kappa \approx 0.35$), which is also included in the computation of the multirater-κ. Hence, κ dropped from approximately 0.6 to 0.45. As we focus on the automatic evaluation and the performance of the automatic system in this article, it is also valid to regard only the reference that was actually shown to the classifier. In this manner we simulate a single rater. Then, κ values are comparable to the semiautomatic, single-rater case ($\kappa \approx 0.6$), i.e., in both cases the classifiers do their task and learn the shown reference in a comparable manner.

The evaluation on a speaker level also had a high and significant correlation of 0.81 ($p<0.01$). The human listeners had an inter-rater correlation of 0.80, which is enough to quantify speech disorders on a speaker level sufficiently. There is no significant difference between human-human and the human-machine correlations ($p>0.05$). Hence, the evaluation of the fully automatic system is at an expert's level. The intrarater correlation of the automatic system is 1 since it always quantifies the same input with the same degree of nasal emissions. The automatic system can be regarded as a fast and reliable way to evaluate nasal emissions in speech of children with CLP at an expert's level. Of course, the application on other phonetic disorders will be realized. Hence, the fully automatic system is suitable for clinical use.

VIII. SUMMARY

This paper presents the first automatic evaluation system for distinct articulation disorders in connected speech. The system has been evaluated on articulation disorders of children with CLP with different extent and characteristics of phonetic disorders. Since the usually applied perceptual evaluation of these disorders requires a lot of time and manpower, there is a need for quick and objective automatic

evaluation. To investigate the evaluation by an automatic system, two experiments with articulation disorders were conducted. On one data set (CLP-1), a test for five characteristic articulation disorders was performed by an experienced speech therapist to show to show that the system is able to detect different articulation disorders. On the second database (CLP-2), the evaluation was performed with a fully automatic system without any additional human effort.

On the frame, phoneme, and word levels, the performance is moderate. On the speaker level, however, the system shows good correlations to the commonly used perceptual evaluation by expert listeners. The correlation between the system and the perceptual evaluation was in the same range as the inter-rater correlation of experienced speech therapists. Thus, the system will facilitate the clinical and scientific evaluation of speech disorders.

ACKNOWLEDGMENTS

This work was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft DFG) under Grant No. SCHU2320/1-1. We thank Dr. Ulrike Wohlleben, Andrea Schädel, and Dorothee Großmann for the expert's annotation of the data. Furthermore, we would like to thank the anonymous reviewers of this article and the editor for their through comments on our work.

[1]R. Ruben, "Redefining the survival of the fittest: Communication disorders in the 21st century," Laryngoscope **110**, 241–245 (2000).
[2]B. Eppley, J. van Aalst, A. Robey, R. Havlik, and M. Sadove, "The spectrum of orofacial clefting," Plast. Reconstr. Surg. **115**, 101e–114e (2005).
[3]M. Tolarova and J. Cervenka, "Classification and birth prevalence of orofacial clefts," Am. J. Med. Genet. **75**, 126–137 (1998).
[4]H. Kawamoto, "Rare craniofacial clefts," in *Plastic Surgery*, edited by J. C. McCarthy (Saunders, Philadelphia, PA, 1990), Vol. **4**.
[5]D. Sell, P. Grunwell, S. Mildinhall, T. Murphy, T. Cornish, D. Bearn, W. Shaw, J. Murray, A. Williams, and J. Sandy, "Cleft lip and palate care in the United Kingdom—The Clinical Standards Advisory Group (CSAG) study. Part 3: Speech outcomes," Cleft Palate Craniofac J. **38**, 30–37 (2001).
[6]J. Karling, O. Larson, R. Leanderson, and G. Henningsson, "Speech in unilateral and bilateral cleft palate patients from Stockholm," Cleft Palate Craniofac J. **30**, 73–77 (1993).
[7]K. Van Lierde, M. D. Bodt, J. V. Borsel, F. Wuyts, and P. V. Cauwenberge, "Effect of cleft type on overall speech intelligibility and resonance," Folia Phoniatr Logop **54**, 158–168 (2002).
[8]K. Van Lierde, M. D. Bodt, I. Baetens, V. Schrauwen, and P. V. Cauwenberge, "Outcome of treatment regarding articulation, resonance and voice in Flemish adults with unilateral and bilateral cleft palate," Folia Phoniatr Logop **55**, 80–90 (2003).
[9]M. Hardin, D.-R. Van Demark, H. Morris, and M. Payne, "Correspondence between nasalance scores and listener judgments of hypernasality and hyponasality," Cleft Palate Craniofac J. **29**, 346–351 (1992).
[10]T. Pruthi and C. Y. Espy-Wilson, "Automatic classification of nasals and semivowels," in *ICPhS 2003-15th International Congress of Phonetic Sciences*, Barcelona, Spain (2003), pp. 3061–3064.
[11]T. Pruthi and C. Y. Espy-Wilson, "Acoustic parameters for automatic detection of nasal manner," Speech Commun. **43**, 225–239 (2004).
[12]T. Pruthi, C. Y. Espy-Wilson, and H. Brad, "Story, simulation and analysis of nasalized vowels based on magnetic resonance imaging data," J. Acoust. Soc. Am. **121**, 3858–3873 (2007).
[13]D. Cairns, J. Hansen, and J. Kaiser, "Recent advances in hypernasal speech detection using the nonlinear teager energy operator," in *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)* (ISCA, Philadelphia, PA, 1996), Vol. **2**, pp. 780–783.
[14]R. Kataoka, K. Michi, K. Okabe, T. Miura, and H. Yoshida, "Spectral properties and quantitative evaluation of hypernasality in vowels," Cleft Palate Craniofac J. **33**, 43–50 (1996).
[15]B. Atal and M. Schroeder, "Predictive coding of speech signals," in *Proceedings of the Conference Communication and Processing* (1967), pp. 360–361.
[16]G. Fant, "Nasal sounds and nasalization," *Acoustic Theory of Speech Production* (Mouton, The Hague, The Netherlands, 1960).
[17]A. Zečević, "Ein sprachgestütztes Trainingssystem zur Evaluierung der Nasalität (A speech-supported training system for the evaluation of nasality)," Ph.D. thesis, University of Mannheim, Mannheim, Germany (2002).
[18]D. Cairns, J. Hansen, and J. Riski, "A noninvasive technique for detecting hypernasal speech using a nonlinear operator," IEEE Trans. Biomed. Eng. **43**, 35–45 (1996).
[19]K. Keuning, G. Wieneke, and P. Dejonckere, "The intrajudge reliability of the perceptual rating of cleft palate speech before and after pharyngeal flap surgery: The effect of judges and speech samples," Cleft Palate Craniofac J. **36**, 328–333 (1999).
[20]S. Paal, U. Reulbach, K. Strobel-Schwarthoff, E. Nkenke, and M. Schuster, "Evaluation of speech disorders in children with cleft lip and palate," J. Orofac. Orthop. **66**, 270–278 (2005).
[21]F. Wuyts, M. D. Bodt, G. Molenberghs, M. Remacle, L. Heylen, B. Millet, K. V. Lierde, J. Raes, and P. V. Heyning, "The dysphonia severity index: An objective measure of vocal quality based on a multiparameter approach," J. Speech Lang. Hear. Res. **43**, 796–809 (2000).
[22]A. Maier, C. Hacker, E. Nöth, E. Nkenke, T. Haderlein, F. Rosanowski, and M. Schuster, "Intelligibility of children with cleft lip and palate: Evaluation by speech recognition techniques," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, Hong Kong, China (2006), Vol. **4**, pp. 274–277.
[23]A. Maier, F. Hönig, C. Hacker, M. Schuster, and E. Nöth, "Automatic evaluation of characteristic speech disorders in children with cleft lip and palate," in Interspeech 2008-*Proceedings of the International Conference on Spoken Language Processing*, 11th International Conference on Spoken Language Processing, Brisbane, Australia (2008), pp. 1757–1760.
[24]A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "PEAKS—A system for the automatic evaluation of voice and speech disorders," Speech Commun. **51**, 425–437 (2009).
[25]A. Fox, "PLAKSS—Psycholinguistische Analyse kindlicher Sprechstörungen (Psycholinguistic analysis of children's speech disorders)," Swets and Zeitlinger, Frankfurt a.M., Germany, now available from Harcourt Test Services GmbH, Germany (2002).
[26]J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Atlanta, GA (1996), Vol. **1**, pp. 349–352.
[27]T. Bocklet, "Optimization of a speech recognizer for medical studies on children in preschool and primary school age," Diplomarbeit, Chair of Pattern Recognition, University Erlangen-Nuremberg, Erlangen, Germany (2007).
[28]A. Maier, "Recognizer adaptation by acoustic model interpolation on a small training set from the target domain," Diplomarbeit, Chair of Pattern Recognition, University Erlangen-Nuremberg, Erlangen, Germany (2005).
[29]A. Maier, T. Haderlein, and E. Nöth, "Environmental adaptation with a small data set of the target domain," in *Lecture Notes in Artificial Intelligence, Ninth International Conference on Text, Speech and Dialogue (TSD)*, edited by P. Sojka, I. Kopeček, and K. Pala (Springer, Berlin, 2006), Vol. **4188**, pp. 431–437.
[30]A. Maier, *Speech Recognizer Adaptation* (VDM, Saarbrücken, 2008).
[31]*Time Warps, String Edits, and Macromolecules*, edited by D. Sankoff and J. Kruskal (Addison-Wesley, Reading, MA, 1983).
[32]F. Gallwitz, *Integrated Stochastic Models for Spontaneous Speech Recognition* (Logos, Berlin, 2002), Vol. **6**.
[33]J. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser output voting error reduction," in *Proceedings of the IEEE ASRU Workshop*, Santa Barbara, CA (1997), pp. 347–352.
[34]A. Maier, C. Hacker, S. Steidl, E. Nöth, and H. Niemann, "Robust parallel speech recognition in multiple energy bands," in *Lecture Notes in Computer Science, Pattern Recognition, 27th DAGM Symposium*, Vienna, Austria, edited by G. Kropatsch, R. Sablatnig, and A. Hanbury (Springer, Berlin, 2005), Vol. **3663**, pp. 133–140.
[35]A. Maier, *Parallel Robust Speech Recognition* (VDM, Saarbrücken, 2008).
[36]H. Niemann, *Klassifikation von Mustern (Pattern Classification)*, 2nd ed. (Springer, Berlin, 2003), http://www5.informatik.uni-erlangen.de/Personen/niemann/klassifikation-von-mustern/m00links.html (Last viewed 02/12/2008).

[37]T. Cincarek, "Pronunciation scoring for non-native speech," Diplomarbeit, Chair of Pattern Recognition, University Erlangen-Nuremberg, Erlangen, Germany (2004).
[38]C. Hacker, T. Cincarek, A. Maier, A. Heßler, and E. Nöth, "Boosting of prosodic and pronunciation features to detect mispronunciations of non-native children," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (IEEE Computer Society Press, Mayi, HI, 2007), Vol. **4**, pp. 197–200.
[39]T. Haderlein, E. Nöth, M. Schuster, U. Eysholdt, and F. Rosanowski, "Evaluation of tracheoesophageal substitute voices using prosodic features," in *Proceedings of the Speech Prosody, Third International Conference*, edited by R. Hoffmann and H. Mixdorff (TUD-Press, Dresden, 2006), pp. 701–704.
[40]T. Haderlein, D. Zorn, S. Steidl, E. Nöth, M. Shozakai, and M. Schuster, "Visualization of voice disorders using the Sammon transform," in *Lecture Notes in Artificial Intelligence, Ninth International Conference on Text, Speech and Dialogue (TSD)*, edited by P. Sojka, I. Kopeček, and K. Pala (Springer, Berlin, 2006), Vol. **4188**, pp. 589–596.
[41]C. Snoek, M. Worring, and A. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th Annual ACM International Conference of Multimedia* (ACM, New York, 2005), pp. 399–402.
[42]S. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, Signal Process. **28**, 357–366 (1980).
[43]M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F. Rosanowski, U. Eysholdt, and E. Nöth, "Evaluation of speech intelligibility for children with cleft lip and palate by automatic speech recognition," Int. J. Pediatr. Otorhinolaryngol. **70**, 1741–1747 (2006).
[44]K. Riedhammer, G. Stemmer, T. Haderlein, M. Schuster, F. Rosanowski, E. Nöth, and A. Maier, "Towards robust automatic evaluation of pathologic telephone speech," in *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)* (IEEE Computer Society Press, Kyoto, Japan, 2007), pp. 717–722.
[45]A. Maier, T. Haderlein, F. Stelzle, E. N. E. Nkenke, F. Rosanowski, A. Schützenberger, and M. Schuster, "Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer," EURASIP J. Audio, Speech, and Music Processing **2010**, In press.
[46]A. Maier, E. Nöth, A. Batliner, E. Nkenke, and M. Schuster, "Fully automatic assessment of speech of children with cleft lip and palate," Informatica **30**, 477–482 (2006).
[47]J. Sammon, "A nonlinear mapping for data structure analysis," IEEE Trans. Comput. **C-18**, 401–409 (1969).
[48]P. Mahalanobis, "On the generalised distance in statistics," in Proceedings of the National Institute of Science of India **12**, 49–55 (1936).
[49]M. Shozakai and G. Nagino, "Analysis of speaking styles by two-dimensional visualization of aggregate of acoustic models," in *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)* (ISCA, Jeju Island, Korea, 2004), Vol. **1**, pp. 717–720.
[50]W. Naylor and B. Chapman, "WNLIB homepage," http://www.willnaylor.com/wnlib.html (Last viewed 07/20/2007).
[51]M. Nagino and G. Shozakai, "Building an effective corpus by using acoustic space visualization (cosmos) method," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (IEEE Computer Society Press, Philadelphia, PA, 2005), pp. 449–452.
[52]A. Maier, M. Schuster, U. Eysholdt, T. Haderlein, T. Cincarek, S. Steidl, A. Batliner, S. Wenhardt, and E. Nöth, "QMOS—A robust visualization method for speaker dependencies with different microphones," J. Pattern Recognition Research **4**, 32–51 (2009).
[53]A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke, "The prosody module," in *Verbmobil: Foundations of Speech-to-Speech Translation*, edited by W. Wahlster (Springer, New York, 2000), pp. 106–121.
[54]P. Bagshaw, S. Hiller, and M. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching," in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)* (ISCA, Berlin, 1993), pp. 1003–1006.
[55]R. Kompe, *Prosody in Speech Understanding Systems*, Lecture Notes in Artificial Intelligence Vol. **1307** (Springer, Berlin, 1997).
[56]S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech* (Logos, Berlin, 2009).
[57]C. Hacker, T. Cincarek, R. Gruhn, S. Steidl, E. Nöth, and H. Niemann, "Pronunciation feature extraction," in *Lecture Notes in Computer Science, Pattern Recognition, 27th DAGM Symposium*, Vienna, Austria, edited by G. Kropatsch, R. Sablatnig, and A. Hanbury (Springer, Berlin, 2005), Vol. **3663**, pp. 141–148.
[58]T. Cincarek, R. Gruhn, C. Hacker, E. Nöth, and S. Nakamura, "Gaikokugohatsuon no jidouhyoutei to yomiayamatta tango no jidoukenshutsu (Automatic evaluation of foreign language pronunciation and automatic recognition of reading errors in vocabulary)," in *Proceedings of the Acoustical Society of Japan* (2004), pp. 165–166.
[59]T. Cincarek, R. Gruhn, C. Hacker, E. Nöth, and S. Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-native's first language," Comput. Speech Lang. **23**, 65–99 (2009).
[60]S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," Speech Commun. **30**, 95–108 (2000).
[61]H. Teager and S. Teager, "Evidence for nonlinear production mechanisms in the vocal tract," in *Speech Production and Speech Modelling* (1990), pp. 241–261.
[62]A. Reuß, "Analysis of speech disorders in children with cleft lip and palate on phoneme and word level," Studienarbeit, Chair of Pattern Recognition, University Erlangen-Nuremberg, Erlangen, Germany (2007).
[63]I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. (Kaufmann, San Fransisco, CA, 2005).
[64]R. Holte, "Very simple classification rules perform well on most commonly used datasets," Mach. Learn. **11**, 63–91 (1993).
[65]E. Frank, Y. Wang, S. Inglis, G. Holmes, and I. Witten, "Using model trees for classification," Mach. Learn. **32**, 63–76 (1998).
[66]G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *11th Conference on Uncertainty in Artificial Intelligence* (Kaufmann, San Mateo, CA, 1995), pp. 338–345.
[67]R. Quinlan, *C4.5: Programs for Machine Learning* (Kaufmann, San Mateo, CA, 1993).
[68]E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," in *15th International Conference on Machine Learning*, edited by J. Shavlik (Kaufmann, San Mateo, CA, 1998), pp. 144–151.
[69]L. Breiman, "Random forests," Mach. Learn. **45**, 5–32 (2001).
[70]B. Schölkopf, "Support vector learning," Ph.D. thesis, Technische Universität Berlin, Berlin, Germany (1997).
[71]Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *13th International Conference on Machine Learning* (Kaufmann, San Mateo, CA, 1996), pp. 148–156.
[72]M. Davies and J. Fleiss, "Measuring agreement for multinomial data," Biometrics **38**, 1047–1051 (1982).
[73]K. Pearson, "Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia," Philos. Trans. R. Soc. London **187**, 253–318 (1896).
[74]D. Seppi, A. Batliner, B. Schuller, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, and V. Aharonson, "Patterns, prototypes, performance: Classifying emotional user states," in Interspeech 2008-*Proceedings of the International Conference on Spoken Language Processing*, 11th International Conference on Spoken Language Processing, Brisbane, Australia (2008), pp. 601–604.
[75]A. Neri, C. Cuchiarini, and C. Strik, "Feedback in computer assisted pronunciation training: Technology push or demand pull?," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (IEEE Computer Society Press, Orlando, FL, 2002), pp. 1209–1212.