Improving the recognition of pathological voice using the discriminant HLDA transformation

Othman Lachhab, Joseph Di Martino, El Hassane Ibn Elhaj, Ahmed Hammouch

Improving the recognition of pathological voice using the discriminant HLDA transformation

Othman LACHHAB
ENSET
Mohammed V University
Rabat, MOROCCO
othmanlachhab@yahoo.fr

Joseph Di MARTINO
LORIA
University of Lorraine
Vandœuvre-lès-Nancy, FRANCE
jdm@loria.fr

El Hassane Ibn ELHAJ
INPT
Madinat Al Irfane
Rabat, MOROCCO
ibnelhaj@inpt.ac.ma

Ahmed HAMMOUCH
ENSET
Mohammed V University
Rabat, MOROCCO
hammouch_a@yahoo.com

***Abstract*— In this paper, we propose a simple and fast method for evaluating the pathological voice (esophageal) by applying the continuous speech recognition in a speaker dependent mode, on our own database of the pathological voice, we call FPSD (French Pathological Speech Database). The recognition system used is implemented using the HTK platform, based on HMM/GMM monophone models. The acoustic vectors are linearly transformed by the HLDA (Heteroscedastic Linear Discriminant Analysis) method to reduce their size in a smaller space with good discriminative properties. The obtained phone recognition rate (63.59 %) is very promising when we know that esophageal voice contains unnatural sounds, difficult to understand.**

Keywords—Automatic Speech Recognition(ASR); HMM; HTK; Pathological voices; HLDA; GMM; MFCC

I. INTRODUCTION

Esophageal voice is a substitution voice learned by a laryngectomee, i.e. a person who has had his or her vocal folds removed, after total laryngectomy. This voice contains specific noises which make it difficult to understand. This noisy speech is produced by a sound generation process which consists in inhaling air into the pharynx and releasing it through the esophagus. The envelope of the waveform and spectral components of esophageal speech do not vary as well as those of laryngeal speech. Furthermore, the pitch of esophageal speech is lower and less stable than the pitch of laryngeal speech. Therefore, the analysis and extraction of F0 processes fail. All these characteristics of esophageal speech cause a production of unnatural sounds.

The recognition and evaluation of pathological voice (esophageal), is a sensitive subject, and the focus of many studies in the field of biomedical applications of speech technology [1] [2]. Esophageal speech can be assessed either by perception judgments or objective analysis techniques. The first type of analysis is the essential method used in clinical practice. It consists in qualifying the pathological voice quality by carefully listening to the patient. However, this method suffers from several drawbacks. Firstly, the perceptual judgment must be made by a jury of experts in order to increase its reliability. Secondly, this perceptual analysis is expensive in time and human resources and cannot be planned easily. On the contrary, the objective analysis [3] [4] is increasingly used. It is based on the analysis of acoustic, aerodynamic and physiological measures. These measures can be directly extracted from the speech signal using a computer system. This objective approach provides acceptable results, but is still insufficient to assess esophageal voice.

The objective of this work is to propose a simple and fast method for evaluating the pathological voice (esophageal) by applying a continuous speech recognition system on our own FPSD (French Pathological Speech Database) database. Our recognition system designed for this specific task is implemented using the HTK platform [5], based on HMM/GMM monophone models. The feature vectors are linearly transformed by the HLDA (Heteroscedastic Linear Discriminant Analysis) method [6] to reduce their size in a smaller space which preserves the discriminant information.

This paper is organized as follows: section 2 details how the previous and current works have attempted to improve the quality/recognition process of the pathological voices. The corpus used concerning the pathological voice, the HLDA transformation method and the recognition system are described in sections 3, 4 and 5 respectively. In section 6, we present the conducted experiments and the obtained results. Finally, section 7 provides a conclusion of this work.

II. PREVIOUS AND CURRENT RESEARCH ON ENHANCING PATHOLOGICAL SPEECH

The esophageal voice is a pathological voice which is difficult to understand because it is too far from a normal laryngeal voice. For this reason, several approaches for improving the quality of pathological voices have been attempted. One of them is whisper speech enhancement by a Code-Excited Linear Prediction (CELP) [14]. This approach uses an external prosthesis in order to allow a natural output voice. However, as it is difficult to generate realistic excitation signals, the resynthesized speech by such a method sounds artificial. Other attempts to enhance the pathological voices, based on the modification of their acoustic features, e.g. by using comb filtering [15], auditory masking [16], and formant synthesis [17], have been proposed. Although these techniques are useful to improve the quality of pathological speech, it is in practice difficult for them to compensate for the acoustic feature differences between pathological and laryngeal speech.

Recently, alaryngeal speech enhancement based on statistical conversion methods has been proposed in [18]. This approach consists in using two parallel corpora, one related to the source (alaryngeal) voice and the other to the target (laryngeal) voice. These corpora contain the same phonetic information. A statistical transformation function is then calculated for converting the source speech signal, in a manner to be perceived as pronounced by the target speaker. Furthermore in order to control the quality of the converted speech, a new conversion method based on Eigen voice conversion (EVC) has been proposed. These conversion processes allow removing the specific noise and improving the intelligibility and quality of the alaryngeal speech.

III. THE FPSD CORPUS

The corpora of pathological speech are relatively less numerous compared to those related to laryngeal speech. This is why we chose to design our own esophageal speech French database entitled FPSD (French Pathological Speech Database). This acoustic and phonetic database is dedicated to the recognition of pathological speech. It contains the recordings of 480 sentences spoken by one laryngectomee. These sentences are classified into 5 categories:

C1) Sentences with one-syllable words.

C2) Sentences with words of one and two syllables.

C3) Sentences with words of three syllables.

C4) Sentences with falling intonation.

C5) Sentences with rising intonation.

This database was divided into two parts: one for training containing 425 sentences and the other for the test containing 55 sentences.

The file structure of the FPSD database is similar to the one used in the TIMIT corpus [7]. We have for each sentence, a wave file (.wav) sampled at 16 kHz (16 bits) with a single input channel, a file (.txt) containing the French text, a file (.wrd) containing the word transcription and a file (.phn) containing the phonetic segmentation. The manual segmentation we conducted is based on our experience in reading spectrograms, listening to the wave sounds, visual examination of the waveforms, and other parameters such as energy and formants. The labeling of the sentences was carried out using SAMPA [8] (Speech Assessment Methods Phonetic Alphabet) a phonetic alphabet that differs a little bit from the International Phonetic Alphabet (IPA), and which presents the advantage of using only simple ASCII characters. Table 1 gives a list of the 36 French phonetic labels used in our own FPSD database, with the IPA correspondence and examples.

TABLE I. SAMPA TRANSCRIPTION OF THE STANDARD FRENCH PHONES.

IPA	SAMPA	Example	IPA	SAMPA	Example
p	p	**p**ont [po~]	j	j	**i**on [jo~]
b	b	**b**on [bo~]	m	m	**m**ont [mo~]
t	t	**t**emps [ta~]	n	n	**n**om [no~]
d	d	**d**ans [da~]	ŋ	N	ri**ng** [ri**N**]
k	k	**c**oût [ku]	l	l	**l**ong [lo~]
g	g	**g**ant [ga~]	ʁ	R	**r**ond [Ro~]
f	f	**f**emme [fam]	w	w	q**u**oi [kwa]
v	v	**v**ent [va~]	ɥ	H	j**u**in [ZHe~]
s	s	**s**ans [sa~]	i	i	s**i** [si]
z	z	**z**one [zOn]	e	e	bl**é** [ble]
ʃ	S	**ch**amp [Sa~]	ɛ	E	s**ei**ze [sEz]
ʒ	Z	**g**ens [Za~]	a	a	p**a**tte [pat]
ɔ	O	c**o**mme [kOm]	ø	2	d**eu**x [d2]
o	o	gr**o**s [gRo]	œ	9	n**eu**f [n9f]
u	u	d**ou**x [du]	oɛ̃	9~	br**un** [br9~]
y	y	d**u** [dy]	ɛ̃	e~	v**in** [ve~]
ə	@	d**e** [d@]	ã	a~	v**ent** [va~]
sil	- ou sil	**silence**	ɔ̃	o~	b**on** [bo~]

IV. THE HLDA TANSFORMATION

The objective of the discriminant transformation method used, the Heteroscedastic Linear Discriminant Analysis HLDA [6], consists in finding a projection space with a reduced dimension of the acoustic vectors. This projection space determined by the HLDA transform must preserve the discriminant information. This processing is based on a maximum likelihood estimation of a matrix M projecting linearly n-dimensional vectors to p-dimensional vectors with p ≤ n.

$$Y = M_p^T X \qquad (1)$$

In order to obtain the transformed data Y, we multiply the transformation matrix M^T of size (p x n) by the original data X. The iterative maximum likelihood algorithm [9] [12] is used in our experiment to estimate the matrix M.

V. THE PHONE RECOGNITION SYSTEM

A. *Speech processing*

The recognition system uses the Mel-Frequency Cepstral Coefficients MFCC [13] and energy, as well as the differential coefficients of these parameters. The speech signal is sampled at 16 KHz and pre-emphasized with a factor of 0.97. The 12 static cepstral coefficients are calculated from a Hamming window of 25 ms shifted every 10 ms, obtained from a bank of

26 Mel scale filters. The logarithm of the energy of the frame is added to the 12 cepstral coefficients in order to create a vector of 13 coefficients. We included also the differential coefficients of order 1, 2 and 3 called dynamic coefficients (Δ, ΔΔ and ΔΔΔ) automatically using the parameterization of the HTK tool. So the acoustical vectors used have at most d = 52 coefficients. Then the space dimension is reduced by the HLDA method applied on all the vectors (training and test) in order to obtain relevant and discriminant vectors with 39 coefficients (d = 39) which represents the reference dimensionality used in most Automatic Speech Recognition (ASR) systems.

B. Context-independent HMM training

The phone recognition system uses 36 phones described in Section 2 (see Table 1). These phones are represented by left-to-right models HMM with five states (but only three of them are emitting observations). Fig. 1 illustrates the topology and the type of HMM used. The training of the models is the starting point of any (ASR) system and certainly the most crucial. It consists in determining the optimal parameters $\Theta = \{A, \pi_i, B\}$.

- π_i : An initial state probability.
- $A = a_{ij}$: The probability of transition from state i to state j (A is a transition probability matrix).
- $B = b_i(o_t)$: the matrix containing the distribution probability of emission the observation o_t in state i.

The output distribution $b_i(o_t)$ for observing o_t in state i is generated by a Gaussian Mixture Model (GMM) and more precisely by a mixture of multivariate Gaussian distribution probabilities $\mathcal{N}(o_t,\mu_{ik},\Sigma_{ik})$ of mean vector μ_{ik} and covariance matrix Σ_{ik} :

$$b_i(o_t) = \sum_{k=1}^{n_i} \frac{c_{ik}}{\sqrt{(2\pi)^d |\Sigma_{ik}|}} \exp\left(-\frac{1}{2}(o_t - \mu_{ik})^T \Sigma_{ik}^{-1}(o_t - \mu_{ik})\right) \quad (2)$$

$$\left(\text{with } \sum_{k=1}^{n_i} c_{ik} = 1\right)$$

Where n_i represents the number of Gaussians in state i, o_t corresponds to an observation o at time t and c_{ik} represents the mixture weight for the kth Gaussian in state i. The recognition system is developed using the platform HTK. For each HMM phonetic model, the Hinit tool initializes the emission probabilities of observations and state transitions through the "segmental k-means" iterative process based on the Viterbi algorithm. These parameters are further refined by an estimation of the maximum likelihood criterion MLE [10] calculated by the Baum-Welch algorithm using the HRest tool. The final learning phase consists in re-estimating simultaneously all of the models on continuous speech data using the HERest tool.

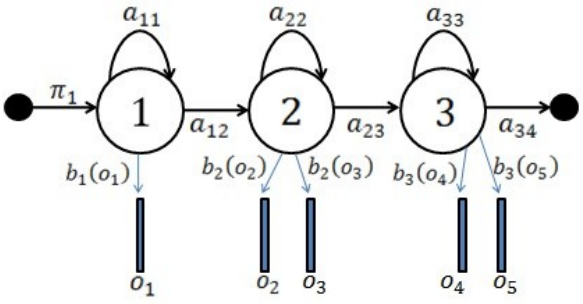

Fig. 1. Topology of the context-independent phonetic HMM.

It is important to choose the appropriate number of Gaussians associated to each state, by making a compromise between a good modeling of the phonetic HMM units and the limited number of training data. A too high number of Gaussians compared to the amount of available data leads to poor learning, because the training data has a limited number of samples for each phoneme. In our case we used 16 Gaussians for each state except for the phoneme /N, which was used with at most 14 Gaussians for each state because of the small available data concerning this phoneme.

C. Phone recognition

The continuous speech recognition is a delicate process because we do not know the boundaries of the phones in the test sentences. In addition monophone HMM models assume that speech is produced as a concatenation of phones which are not affected by the phonetic contexts neighbors left/right and right/left (context-independent). In order to perform the recognition process, it is useful to determine the sequence of states that has generated the given observations. In fact, from the sequence of states, we can easily find the most probable phone string that match the parameters observed. This task is performed by the Viterbi decoding algorithm applied on each test sentences using the optimal parameters $\{A, \pi_i, B\}$ already estimated. A bigram language model is calculated on all of the training data to improve the decoding.

VI. EXPERIMENTS AND RESULTS

In order to evaluate the recognition of phones, we chose to perform our first tests on the TIMIT [7] database built from laryngeal read speech. We used the same labeling of 39 phonetic classes described by K. F. Lee and H. W. Hon [11]. We conducted four experiments with the HTK recognition system to evaluate the contribution of differential coefficients and the HLDA transformation. In the first experiment we worked with acoustic vectors of dimension d = 39 (12 MFCC, E; 12 ΔMFCC, ΔE; 12 ΔΔ MFCC, ΔΔE) representing the reference dimensionality in most ASR systems. For the second experiment the derivative of order 3 (ΔΔΔ) is included in the space vectors to increase the number of coefficients to d = 52 (12 MFCC, E; 12 ΔMFCC, ΔE; 12 ΔΔMFCC, ΔΔE; 12 ΔΔΔMFCC, ΔΔΔE). The third experiment consists in applying the discriminant transformation HLDA (39→39) on the 39 feature vectors used in experiment 1 without dimension

reduction. Whilst in the 4th and last experiment the number of 52 coefficients used in experiment 2 is reduced to 39 coefficients by the HLDA (52→39) transformation. We run all these experiments also on our FPSD database. Concerning this database, we use 36 phonetic labels described in section 2 (see Table 1). And for each state 16 Gaussians are used except for the phoneme /N which used only 14 Gaussians for each state. A bigram language model is calculated on the all of the training FPSD database containing 425 sentences, to improve the phone decoding on the esophageal voice.

The recognition rate (Accuracy) of phones is calculated by Eq. 3, where N is the total number of labels of the test utterances and S, I and D (resp.) are the Substitution, Insertion and Deletion errors, computed by the DTW algorithm (Dynamic Time Warping) between the correct phone string and the recognized phone string.

$$Accuracy = \frac{N-(S+D+I)}{N};\quad Correct = \frac{N-(S+D)}{N} \qquad (3)$$

The two tables, Table 2 and Table 3 present the phone recognition rates for the four experiments described above, respectively on the core test of the laryngeal voice TIMIT database, and the test part of our own FPSD database.

We note that the results observed in experiment 4 provide a significant improvement in phone recognition rate (accuracy) compared to the other experiments.

TABLE II. INFLUENCE OF THE NUMBER OF DIFFERENTIAL COEFFICIENTS WITH THE HLDA TRANSFORMATION ON PHONE RECOGNITION RATES ON THE CORE TEST OF THE TIMIT DATABASE.

39 monophone HMMs with 16 Gaussians per state+ Bigram	**Accuracy (%)**	**Correct (%)**
Exp 1 : 39 MFCC coefficients	69.19	71.78
Exp 2 : 52 MFCC coefficients	67.96	71.38
Exp 3 : HLDA (39 →39)	70.24	72.77
Exp 4 : HLDA (52 →39)	**71.32**	**74.07**

TABLE III. INFLUENCE OF THE NUMBER OF DIFFERENTIAL COEFFICIENTS WITH THE HLDA TRANSFORMATION ON PHONE RECOGNITION RATES ON THE TEST PART OF OUR OWN FPSD DATABASE.

36 monophone HMMs with 16 Gaussians per state + Bigram	**Accuracy (%)**	**Correct (%)**
Exp 1 : 39 MFCC coefficients	61.89	67.62
Exp 2 : 52 MFCC coefficients	58.49	65.29
Exp 3 : HLDA (39 →39)	62.31	66.88
Exp 4 : HLDA (52 →39)	**63.59**	**69.43**

VII. CONCLUSION

This paper, contributes to the continuous speech recognition of the pathological voice. Our ASR system based on context-independent HMM/GMM models (monophone) exhibits a significant improvement in phone recognition accuracy (63.59 %) by using the discriminant HLDA transformation and high order differential coefficients. These results are encouraging. Indeed, the performance of our system can be further improved by extending our FPSD corpus in order to use context-dependent HMM models (triphones).

REFERENCES

[1] D. Pravena, S. Dhivya, A. Durga Devi, "Pathological Voice Recognition for Vocal Fold Disease", International Journal of Computer Applications, Vol-47, No. 13, 2012.

[2] A. A. Dibazar, T. W. Berger, and S. Narayanan, "Pathological voice assessment," Engineering in Medicine and Biology Society, 2006, EMBS '06, 28th Annual International Conference of the IEEE, pp. 1669-1673, August 2006.

[3] L. Wuyts, M. S. De Bodt, G. Molenberghs, M. Remacle, L. Heylen, B. Millet, K. Van Lierde, J. Raes, and P. H. Van de Heyning, "The dysphonia severity index : an objective measure of vocal quality based on a multiparameter approach", In Journal of Speech, Language, and Hearing Research 43, pages 796–809, 2000.

[4] P. Yu, M. Ouakine, J. Revis, and A. Giovanni, "Objective voice analysis for dysphonic patients: a multiparametric protocol including acoustic and aerodynamic measurements", In Journal Voice 15 , pages 529–542, 2001.

[5] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland, "The HTK Book Revised for HTK Version 3.4", December 2006.

[6] N. Kumar and A. Andreou, "Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition", Speech Communication, 26(4) : 283–297, 1998.

[7] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D.S. Pallett, N. L. Dahlgren, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM", NTIS order number PB91-100354, 1993.

[8] Speech Assessment Methods Phonetic Alphabet (SAMPA) http://www.phon.ucl.ac.uk/home/sampa/french.htm

[9] M. J. F. Gales "Semi-tied covariance matrices for hidden markov models", IEEE Transactions on Speech and Audio Processing, 7(3) :272–281, 1999

[10] L. R. Rabiner, "Tutorial on hidden Markov models and selected applications in speech recognition", Proc. IEEE, vol. 77, no. 2, pp. 257-278, February 1989.

[11] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," IEEE Trans. ASSP, vol. 37(11), pp. 164–1648, November 1989.

[12] L. Burget, "Combination of speech features using smoothed heteroscedastic linear discriminant analysis", In 8th International Conference on Spoken Language Processing, Jeju island, KR, October 2004.

[13] A. M. Noll, "A Cepstrum speech determination", Journal of the Acoustic Society of America 41 (1), 293-309. 1967.

[14] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec,", IEEE Trans. Biomed. Eng., vol. 57, no. 10, pp. 2448– 2458, October 2010.

[15] A. Hisada and H. Sawada, "Real-time clarification of esophageal speech using a comb filter," International Conference on Disability, Virtual Reality and Associated Technologies, pp. 39–46, 2002.

[16] H. Liu, Q. Zhao, M. Wan, and S. Wang, "Enhancement of electrolarynx speech based on auditory masking", IEEE Trans. Biomed. Eng., vol. 53, no. 5, pp. 865–874, May 2006.

[17] K. Matui, N. Hara, N. Kobayashi, and H. Hirose, "Enhancement of esophageal speech using formant synthesis," Proc. ICASSP, pp. 1831–1834, Phoenix, Arizona, May 1999.

[18] D. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "Alaryngeal speech enhancement based on one-to-many eigenvoice conversion," IEEE Trans. Audio. Speech Language, vol. 22, no.1, pp. 172–183, January 2014.