Pattern Recognition Letters 34 (2013) 335–343

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Aging speech recognition with speaker adaptation techniques: Study on medium vocabulary continuous Bengali speech

Biswajit Das *, Sandipan Mandal, Pabitra Mitra, Anupam Basu

Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur 721302, West Bengal, India

A R T I C L E I N F O

Article history:
Received 23 May 2012
Available online 16 November 2012

Communicated by S. Sarkar

Keywords:
Aging speech recognition
Vocal tract length normalization (VTLN)
Maximum likelihood linear transform (MLLT)
Maximum likelihood linear regression (MLLR)
Maximum a posteriori (MAP)
Maximum mutual information estimation (MMIE)

A B S T R A C T

The article describes the speech recognition system development in Bengali language for aging population with various adaptation techniques. Variability in acoustic characteristics among different speakers degrades speech recognition accuracy. In general, perceptual as well as acoustical variations exists among speakers, but variations are more pronounced between young and aged population. Deviation in voice source features between two age groups, affect the speech recognition performance. Existing automatic speech recognition algorithms demands large amount of training data with all variability to develop a robust speech recognition system. However, speaker normalization and adaptation techniques attempts to reduce inter-speaker or intra-speaker acoustic variability without having large amount of training data. Here, conventional acoustic model adaptation method e.g. vocal tract length normalization, maximum likelihood linear regression and/or maximum a posteriori are combined in the current study to improve recognition accuracy. Moreover, maximum mutual information estimation technique has been implemented in this study.

1. Introduction

Automatic speech recognition (ASR) system having high recognition accuracy have potential of being an useful part of our daily life. It is a well established fact that ASR system performs satisfactorily under controlled condition. However, due to mismatch in acoustic properties of training and testing data, ASR performance degrades rapidly in uncontrolled environment. Among different mismatch in training and testing speech data, aging contribute sufficient deviation in spectral parameters. ASR system that had been developed in Bengali with the training data collected from young population performs well for test data of young people but recognition accuracy degrades drastically for test data of aged population. Due to this observation, it has been our endeavour to develop a robust ASR system for aged population. In this work, speaker adaptive acoustic modeling methods are investigated to compensate acoustic mismatch between training and testing observations due to aging.

Human articulatory system evolves during our life time. Physiological and anatomical changes are studied in several studies (Ulatowska, 1985; Lindblom, 1971; Linville and Rens, 2001; Rother et al., 2002; Xue and Hao, 2003; Paulsen and Tillmann, 1998; Rodeo et al., 1993; Wilcox and Horii, 1980; Yumoto et al., 1984; Krom, 1993; Hillenbrand et al., 1994; Tolep et al., 1995) Physiological and anatomical changes of vocal tract affect different voice source features e.g. fundamental frequency (F_0), formant frequencies ($F_1, F_2, F_3, \ldots$ etc.), *jitter*, *shimmer*, *voice-onset-time* and *harmonic-to-noise ratio* of speech signal. In the study Torre and Barlow (2009) and Vipperla et al. (2010), patterns of variations of those voice source feature with aging has been reported. Different parts of vocal tract are directly or indirectly responsible for different phoneme pronunciation. If a particular part of vocal tract is affected due to aging, acoustic characteristics of related phone will be changed. Physiological and anatomical changes that take place in our articulatory system with aging (Linville, 2001) will be discussed next.

Physiological changes of vocal tract not only depends on chronological age, but other factors which can deteriorate the health of articulatory organs; alcohol absorption, smoking habit, food habit, and medical condition (Linville, 2001; Gorham-Rowan and Laures-Gore, 2006). Another important factor for voice quality degradation is hereditary traits of the family.

Inter-speaker or intra-speaker acoustic variations are the major sources of error in ASR (Huang and Lee, 1991; Digalakis et al., 1995; Wilpon et al., 1996). Different sources of acoustic variations are anatomical characteristics such as vocal tract length, dimension of mouth, nasal cavities and speaking style (e.g. accent, dialect and speaking rate). Variability can be reduced with speaker

* Corresponding author. Mobile: +91 9775550915.
E-mail addresses: biswajit.net@gmail.com (B. Das), mandal.sandipan@gmail.com (S. Mandal), pabitra@gmail.com (P. Mitra), anupambas@gmail.com (A. Basu).

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.10.029

normalization or adaptation techniques. Vocal tract length normalization (VTLN) (Lee and Rose, 1996; Wegmann et al., 1996; Eide and Gish, 1996; Welling et al., 1999; Uebel and Woodland, 1999) and maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995; Gales et al., 1996a) are two conventional techniques which are capable to reduce variability, and also reduce word error rate (WER). There are numerous studies have been done on vocal aging, but there are limited number of experiment has been reported about effect of aging on ASR. In (Vipperla et al., 2010; Giuliani et al., 2006), comparison of speech recognition performance of different age groups with VTLN and MLLR techniques have been reported. In (Huda et al., 2009b,a), modification of expectation maximization algorithm has been reported.

Bengali is one of most widely spoken languages in the world with about 230 million speakers across the globe. Although most predominantly spoken in south-east Asia. It ranks 6th[1] in the list of most commonly spoken languages. In this paper, we will analyze the effect of aging on Bengali phoneme recognition with a large vocabulary corpus of aging Bengali speech. Corpus of elderly people is illustrated in Section 2. Furthermore, we have analyzed word recognition accuracy of different acoustic models. Number of normalization and adaptation techniques are employed to compensate mismatch between training and testing data and improve the recognition accuracy. We have proposed a hybrid adaptation method (Combination of VTLN, LDA, MLLR, MAP and MMIE) in the current study. Proposed hybrid model improves recognition accuracy up to 10–12% in average than baseline ASR system.

2. Bengali speech corpus used in our study

Two Bengali speech corpora are used in this study. One of these corpora consists of speech samples of young speakers, and is marked as BENG_YO. Another corpora is consist of speech signal of aged population, and is referred as BENG_OL. We have used same set of text corpora for building both speech corpora. Bengali speech corpus design has been reported in (Das et al., 2011) previously.

In BENG_OL corpus, we have recorded speech signal from 40 *male* and 20 *female* aged speakers having age 60–80 years. Speech data collection from elderly people is a challenging task because older people are unable to record speech for long sessions. Furthermore, most of the aged people do not feel excited to record their voice. Speakers who are suffering from low vision face difficulty to read the text. We have enlarge the font of the text sufficiently so that elderly people can read without difficulty. As voice quality changes in different session of a single speaker due to different mental and physical states, speech data has been recorded in two sessions. All the recording has been done at room environment. Mother tongue of all speakers is Bengali.

We have selected 7500 text sentences for recording. These are sourced from the Bengali News daily *Anandabazar patrika* and Bengali literature. Sentences are selected with optimal text selection procedure, where a process of balanced phoneme and triphone selection in the text corpus is adopted (Mandal et al., 2011). Optimal text selection aims to keep balance in frequency of each phone in the text corpus. Each sentence is recorded with sample frequency 16000 Hz in *mono channel*. Speech signals are encoded with *16 bit Pulse Code Modulation*. *Sony FV-220* microphone and *Emu speech tool* has been used for speech recording. We have maintained 15 cm distance from mouth to microphone for each speakers at the time of voice recording. Each speaker has recorded almost 200 sentences. There are 19500 words in the phonetic dictionary. Total duration of BENG_OL corpus is 12 h. According to place of articulation and manner of articulation, 42 unique phoneme have been considered in this corpus. The Bengali phonetic dictionary has been created initially with grapheme to phoneme conversion procedure on the words. Finally, the dictionary is corrected manually according to pronunciation variation.

BENG_YO corpus of younger adults is created using the same set of sentences. BENG_YO corpus consists of 60 *male* and 30 *female* speakers. Speech recording configurations are same as stated above. Young speakers are considered with age between 20 and 40 years. Although, we have selected 60 speakers (40 male and 20 female) from 90 speakers to balance with speech corpus of elderly in this experiment. This corpus has been used for comparative purpose. Total duration of BENG_YO corpus is 21 h.

In test data set, there are five male and five female speakers of both age groups. Average age of elderly test speakers is 72 and young speakers is 26. Each speaker has recorded 20 sentences. That are phonetically well balanced. Speakers in test data set are not belong to training set.

3. Acoustic speaker adaptation techniques

Speaker adaptive acoustic model reduces variability among different speakers. A number of physiological and anatomical characteristics of our auditory system are the source of spectral variation e.g. vocal tract length differs from speaker to speaker, speaker's age.

Normalization and adaptation can be categorized into two categories. First one is responsible for feature vector normalization and transformation. Cepstral mean normalization, cepstral variance normalization, linear discriminant analysis and maximum likelihood linear transformation are examples of feature space normalization and transformation techniques. Second one is related to model space transformation techniques such as vocal tract length normalization (VTLN), maximum-a posteriori (MAP) and maximum likelihood linear regression (MLLR).

3.1. Cepstral mean and variance normalization

Cepstral mean and variance normalization is performed on static feature (Liu et al., 1993). It is applied at the time of mel frequency cepstral coefficient computation from speech signal. It forces feature vectors to be zero mean and unit variance. These methods reduce the speaker-to-speaker acoustic variation, and additive noise induced from channel and environment.

3.2. Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA) (Haeb-Umbach and Ney, 1992) is the statistical technique used for maximizing separability among different classes. LDA method try to find out linear transformation of feature vector from n-dimensional to m-dimensional ($m < n$) so that inter class statistical distance increases. The implicit assumption is that the rejected sub-vector does not carry any classification information. For Gaussian models, this assumption is equivalent to the assumption that the means and variances of the class distributions are the same for all classes, in the rejected subspace. Furthermore, LDA assumes that the within-class variances are equal for all the classes. It is applied at the time of context independent model parameter optimization. As it reduces feature vector length, decoding process performs faster than without LDA.

[1] 'Bengali language'. http://en.wikipedia.org/wiki/Bengali_language.

3.3. Maximum likelihood linear transform

Maximum likelihood linear transform (MLLT) can also be employed for the purpose of feature decorrelation (Gopinath, 1998). Under this approach, MLLT applies a linear transform to the acoustic features in an attempt to capture the correlation between the feature vector components. MLLT is applied on top of LDA feature vectors. MLLT is also computed from context independent HMM models using training data. It does not require any extra adaptation data. Transformations and model parameters are optimized simultaneously by maximum likelihood criteria on training data. Transformation matrix W is estimated by maximizing the auxiliary function below:

$$Q(M,\hat{M}) = K - \frac{1}{2}\sum_{s=1}^{J}\sum_{t=1}^{T}\gamma_s(t)\left[K_s + log(|\Sigma_s|) + (o_t - \hat{\mu}_s)\hat{\Sigma}_s^{-1}(o_t - \hat{\mu}_s)^T\right] \tag{1}$$

where $\hat{\mu}_s = W\mu_s + b$ and $\hat{\Sigma}_s = W_s\Sigma_s W_s^T$ are transformed mean vector and covariance matrix respectively. M and $\hat{M}$ are current model parameters and re-estimated model parameters respectively. K is the constant dependent only on transition probabilities, and K_s is the normalization constant associated with Gaussian s. o_t is the training data. $\gamma_s(t)$ is the probability of being in state s at time t. Moreover, transformation matrix W will be applied to the context dependent training process for adapting model parameters. Furthermore the transcription matrix W will be applied for decoding process to adapt the test data. MLLT is a speaker independent adaptation method.

3.4. Maximum a posteriori (MAP)

Maximum a posteriori estimation (Gauvain and Lee, 1992; Lee and Gauvain, 1993) is an way to incorporate prior information in the training process. Maximum likelihood (ML) estimation approach gives inaccurate estimate of model parameters for sparse training data. Baum–Welch re-estimation produces ML estimation of model parameters λ as

$$\lambda_{ML} = argmax_{\lambda}P(O|\lambda) \tag{2}$$

where $P(O|\lambda)$ is the likelihood estimation function. In MAP training, it produces maximum-a posteriori estimation as

$$\lambda_{MAP} = argmax_{\lambda}P(O|\lambda)P(\lambda) \tag{3}$$

where $P(\lambda)$ is the prior probability information. For speaker adaptation, $P(\lambda)$ is derived from the speaker independent model.

3.5. Vocal tract length normalization (VTLN)

VTLN is a frequency warping method of the frequency axis of power spectrum (Lee and Rose, 1996; Wegmann et al., 1996; Eide and Gish, 1996). It is important to estimate perfect frequency scaling factor for each speaker or each utterances. Frequency axis is compressed or stretched for individual speaker with a scaling factor α. The conventional method to estimate scaling factor is the grid search method, where scaling factor is selected from discrete set of scaling factor which will increase likelihood of warped data given a set of acoustic model. Vocal tract length normalization is applied for both training and test data. In this work, we have selected a inverse linear frequency warping method which will adapt frequency axis of male-filter bank. We have used context independent HMM model for finding out best warping factor for each spoken sentence. Initial warping factor is chosen 0.8 which will incremented by 0.05 at each training steps, and will stop at 1.2. It will only consider those warping factor which are applied to normalize the feature vector, and yield maximum likelihood

$$\hat{\alpha} = argmax_{\alpha}\left(P(O^{\alpha}|W,\lambda)\right) \tag{4}$$

where $\hat{\alpha}$ is the optimal scaling factor and O^{α} is observation sequences obtained by applying scaling factor α. W is the uttered words. At the time of training, observation feature vectors of each utterances rather than each speaker are normalized with set of scaling factor. Select an optimal warping factor for each utterance, according to Eq. (4). Then, optimally normalized utterances are stored for further training.

Decoding procedure promotes the normalized model parameters for recognizing the unwarped test data. Phoneme or word hypotheses, generated from recognition are applied with normalized model parameters to select optimal warping factor of test data according to Eq. (4). Finally, optimally normalized feature vectors are promoted to recognition process.

3.6. Maximum likelihood linear regression (MLLR)

A baseline acoustic model is necessary for the maximum likelihood linear regression speaker adaptation technique (Leggetter and Woodland, 1995; Gales et al., 1996a). Baseline acoustic model is then adapted with some adaptive training data. Only the mean vectors of each Gaussian mixture model are considered to adapt in this work. In general, training data are assumed as parameterized speech frame vector $(o_1, o_2, \ldots, o_t)$. Adaptation of mean vector has been done with transformation matrix W_j and extended mean vector ζ_j, as

$$\hat{\mu}_j = A\mu_j + b = W_j\zeta_j \tag{5}$$

where extended mean ζ_j is represented as $[\omega\ \mu_1, \ldots, \mu_n]^T$ (ω = 1 to include offset). W_j is the regression matrix which will maximize likelihood of the adaptation data. This regression matrix is iteratively estimated by forward–backward algorithm in such a way to maximize likelihood of adaptation data. Once W_j is created, it will also be employed for the recognition of the test data. In this process only transformation matrix W_j is to be re-estimated iteratively which can be achieved through an auxiliary function as

$$Q(M,\hat{M}) = K - \frac{1}{2}P(O|M)\sum_{s=1}^{J}\sum_{t=1}^{T}\gamma_s(t)\left[K_s + log|\Sigma_s| + \left(o_t - W_s^T\zeta_s\right) \times \Sigma_s^{-1}\left(o_t - W_s^T\zeta_s\right)^T\right] \tag{6}$$

where M and $\hat{M}$ are current model parameters and re-estimated model parameters respectively, and $\gamma_s(t)$ is the probability of being in state s at time t. $P(O|M)$ is the likelihood of the training data.

Equating the derivatives of Q (M,$\hat{M}$) with respect to W_j, and equating to zero will yield maximum value of Q (M,$\hat{M}$). The general form to compute W_j is given by

$$\sum_{t=1}^{T}\gamma_j(t)\Sigma_j^{-1}o_t\zeta_j^T = \sum_{t=1}^{T}\gamma_j(t)\Sigma_j^{-1}W_j^T\zeta_j\zeta_j^T \tag{7}$$

where $\gamma_j(t)$ is the conditional probability of mixture component j at time t and Σ_j^{-1} is also a diagonal standard deviation matrix of mixture component j. ζ_j is the extended mean. Above auxiliary function ensure the maximum likelihood estimation of adaptation data.

From Eq. (7), computation of W_j will be computationally expensive for full co-variance matrix. Computation overload can be reduced considering diagonal covariance matrix (Leggetter and Woodland, 1995; Gales et al., 1996b). Multiple iteration of MLLR performs well at the time of decoding. In this work, we have implemented multiclass regression matrix for better performance. Multiple regression class matrix can outperform single class regression matrix method. In multiclass regression method, it is required to tie transformation matrices across number of Gaussians. For this reason, number of Gaussians are grouped together using regression class tree. In these process, number of Gaussian component will share statistically similar group or class. It is an important decision making method because depending on this method one good regression class will be constructed. Gaussian component of same acoustic phenomena will be transformed with a specified regression matrix.

4. Maximum mutual information estimation (MMIE)

Maximum mutual information estimation (MMIE) (Valtchev et al., 1997) is a discriminative training approach. MMIE is an alternative method of maximum likelihood estimation (MLE) to re-estimate the HMM model parameters. MMIE is computationally very much complex and time consuming than MLE method. In case of MLE, posterior probability maximization is the only aim with provided training data. Model parameters of other classes do not involved in this parameter re-estimation. MMIE attempts to maximize the posterior probability with corresponding training data. Let there are T observation sequence $O_1, O_2, \ldots, O_T$ and corresponding word transcription is W_t. Then, MMIE objective function will be of the form:

$$F(\lambda) = \sum_{t=1}^{T} log \frac{P_\lambda(O_t|\mu_{W_t})P(W_t)}{\sum_{w} P_\lambda(O_t|\mu_W)P(W)} \quad (8)$$

μ_W is the composite model according to word sequence W and $P(W)$ is the probability of those sequence determined by language model. So, it can be assumed as two stage optimization problem. In the first stage, HMM model parameters are adapted to increase the numerator function. Denominator term is responsible for minimizing its value so that overall likelihood can be maximized. Parameter re-estimation methods has been beautifully described in (Valtchev et al., 1997; Bahl et al., 1986; Normandin et al., 1991). We have implemented the lattice based maximum mutual information (MMI) training method where phone and word based lattices are created with HMM model parameters and unigram language model for phoneme and word recognition respectively. In this process, lattices produces all likely hypothesis of a training utterances instead of looking for single likely transcription.

5. Results and discussion: speech recognition

In this work, CMU SPHINX speech recognition toolkit (Lee et al., 1990) has been employed for Bengali speech recognition system. There are three basic module in speech recognition process e.g. feature extraction, acoustic modeling and language modeling.

We have used 39 dimensional mel frequency cepstral coefficients (MFCC) Molau et al., 2001 feature as input to the acoustic modeling.

Trigram language model of phoneme and words are applied for phoneme and word recognition using triphone acoustic model. CMUCLTK toolkit has been used for building those trigram language models.

Fig. 1 presents a schematic of various speaker adaptation techniques employed in our study. Baseline acoustic model parameters are estimated by maximum likelihood (ML) criteria. Speaker normalization and adaptation techniques are explored to the current study for better recognition performance. Vocal tract length normalization (VTLN), linear discriminative analysis (LDA), maximum likelihood linear transformation (MLLT), maximum likelihood lin-

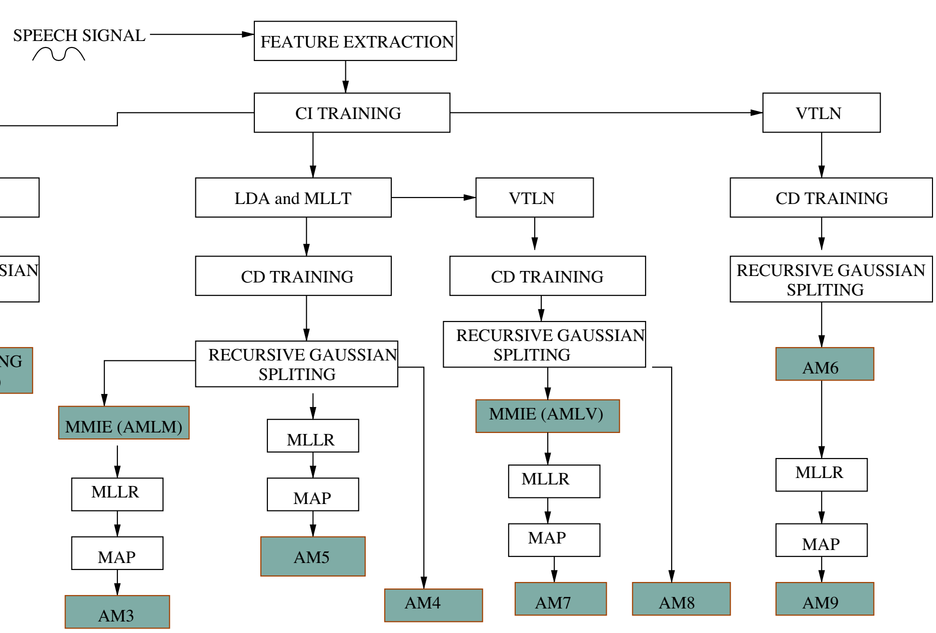

Fig. 1. Incorporation of speaker adaptation techniques in the acoustic models (AM).

Table 1
Phoneme recognition accuracy (%) (standard deviation) of different acoustic models with test data of aged people.

Models	AM1	AM4	AM6	AM8	AM2	AM5	AM9	AMLM	AMLV	AM3	AM7
YO_AMs	51.8 (12.2)	52.7 (12.1)	52.9 (12.2)	53.5 (11.8)	53.5 (12.7)	54.2 (12.6)	55.8 (11.2)	61.2 (11.2)	62.3 (11.3)	63.9 (11.1)	64.4 (11)
OL_AMs	72.6 (11.6)	73.7 (11.7)	73.7 (11.6)	75.5 (8.2)	76.5 (10.6)	78.7 (10.8)	79.8 (10.2)	79.2 (10.1)	80.1 (10)	82.8 (9.5)	83.1 (8)
MIX_AMs	59.5 (18.6)	61.2 (18.9)	62.4 (17.9)	63.9 (17.8)	65.6 (16.2)	68.9 (16.4)	69.4 (15.1)	70.3 (15.4)	72.2 (15.2)	73.8 (15.3)	75.3 (15)

ear regression (MLLR), maximum a posteriori (MAP) and maximum mutual information estimation (MMIE) are applied alone as well as in combination for both age group to compare recognition performance of aged. Combination of adapted acoustic models mentioned in Fig. 1 will be described as follows:

- AM1: AM1 is the baseline system. Its model parameters are estimated through a sequential process. At first, context independent (monophone) HMM model (CIHMMs) parameters are estimated by means of forward–backward algorithm by employing unnormalized training data. We have considered 10 iterations for each stage of model parameter estimation. Then, untied context dependent (triphone) HMM models (CDHMMs) are trained on CIHMMs. A phonetic decision tree built on linguistic questions is used for state tying across different CDHMMs. Finally, the AM1 is obtained by dividing single Gaussian mixture component into eight mixture components by iterative Gaussian splitting method.
- AM2: In acoustic model AM2, we have adapted the AM1 by MLLR and MAP subsequently. Multiple class linear regression matrix is estimated by MLLR technique using adaptation data of a particular speaker. Moreover, model parameters are adapted by MAP adaptation method.
- AM3: It has been achieved after applying MLLR and MAP on maximum mutual information training using adaptation data.
- AM4: This acoustic model is extended from CIHMMs. At first, 29 dimensional optimal MFCC features are extracted from 39 dimensional feature vector by means of LDA. After that speaker independent transformation matrix is optimized by MLLT technique. Rest of the processes are creation of CDHMMs and Gaussian mixture component splitting applying transformation matrix. Same transformation matrix is used at the time of decoding process.
- AM5: In AM5, model parameters of AM4 are adapted by MLLR and MAP techniques as mentioned in AM2 using adaptation data.
- AMLM: In this acoustic model, we have applied MMIE training approach after obtaining acoustic model AM4. At first, phone lattices and word lattices are generated using decoder for phone and word recognition respectively. We have considered unigram phone and word label language model for phone and word lattice generation respectively. Then, pruning of that lattices are conducted by beam pruning method. Finally, model parameters are re-estimated by Baum–Welch algorithm.
- AM6: It is the speaker normalized acoustic model. It has been achieved after applying VTLN technique on CIHMMs.
- AM8: In this AM, we have applied VTLN after speaker LDA and MLLT techniques.
- AM9: We have obtained this AM after applying MLLR and MAP on AM6 using adaptation data.
- AMLV: We have implemented MMIE training on AM8. Procedures are same as discussed in acoustic model AMLM.
- AM7: Acoustic model, AMLV is then adapted with MLLR and MAP subsequently using adaptation data.

In the decoding process, unnormalized, normalized feature and different adapted AMs are exploited for analyzing phoneme and word recognition performance. Test utterances are adapted and/or normalized by transformation matrix, linear regression matrix and VTLN respectively according to various AMs.

5.1. Phoneme recognition with triphone acoustic models

In our pronunciation pattern, current phone is influenced with its left and right phone. In context dependent acoustic model, every triphone is constructed as root phone and its immediate left and right phone like $/A - m + i/$. We have created triphone acoustic model exploiting monophone acoustic models of three different training data. There are three baseline acoustic model, and rest of the triphone models are representative of different adaptive and normalized model. CDHMM models are represented with five state and each state is considered as eight Gaussian mixture components. The results are discussed in the following sections.

5.1.1. Linear discriminative analysis and maximum likelihood linear transform

LDA is applied at the initial stage of training. It selects informative 29 feature entities out of 39 to increase inter-class separability. Moreover, MLLT matrix is estimated through an iterative process using LDA features. MLLT matrix is created on context independent

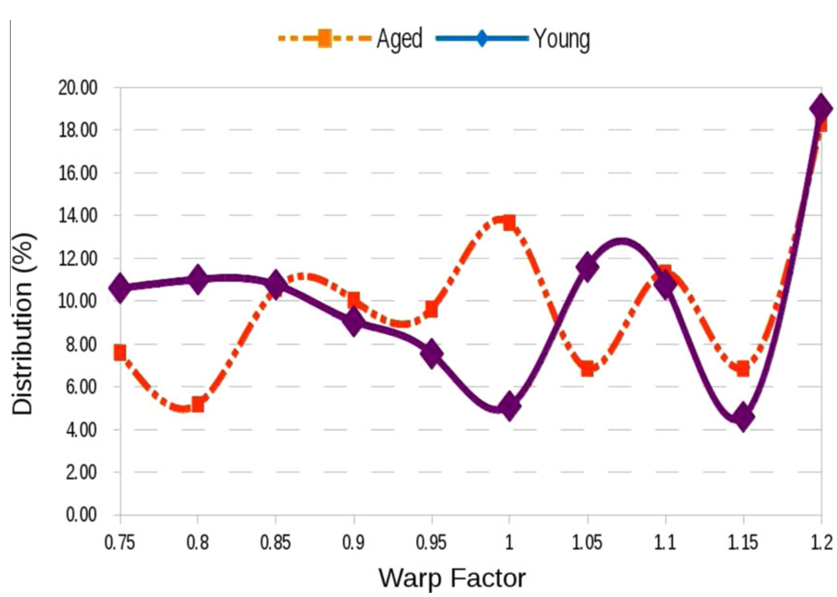

Fig. 2. Distribution of warp factor of young and aged speakers.

Table 2
Confusion matrix of significant vowels and consonants obtained from baseline acoustic model (AM1).

Actual	Predicted												
	A	a	i	o	u	b	j	t	k	n	p	r	T
^*A*	15.26	10.53	–	–	–	–	–	–	–	–	–	–	
a	–	79.5	–	11.5	–	–	–	–	–	–	–	–	
E	13.16	–	–	–	–	–	–	–	–	–	–	–	
e	–	–	5.74	–	–	–	–	–	–	–	–	–	
u	–	–	–	9.02	–	–	–	–	–	–	–	–	
^*o*	–	–	–	16.67	–	–	–	–	–	–	–	–	
^*u*	–	–	–	–	33.33	–	–	–	–	–	–	–	
bh	–	–	–	–	–	14.29	–	–	–	–			
ch	–	–	–	–	–		7.21	8.93					
dh	–	–	–	–	–			3.85		7.69	7.69		
jh	–	–	–	–	–	–	33.3	–	–	–	–	–	–
kh	–	–	–	–	–	–	–	–	15.69	–	–	–	–
m	–	–	–	–	–	2.48	–	–	–	6.67	–	–	–
p	–	–	–	–	–	10.26	–	5.13	–	–	–	–	–
ph	–	–	–	–	–	20	–	–	–	14.29	14.29	–	14.29
R	–	–	–	–	–	–	–	–	–	5.66	–	24.53	–
T	–	–	–	–	–	–	–	13.21	4.08	–	–	7.55	–
Th	–	–	–	–	–	–	–	13.04	8.70	–	–	–	13.04
th	–	–	–	–	–	–	–	15.64	–	–	–	–	–

Table 3
Confusion matrix of significant vowels and consonants obtained from best AMLV acoustic model.

Actual	Predicted													
	A	a	i	o	u	b	j	t	k	n	p	r	th	T
^*A*	5.26	5.26	–	–	–	–	–	–	–	–	–	–	–	
a	–	84.06	–	8.21	–	–	–	–	–	–	–	–	–	
E	11.11	–	–	–	–	–	–	–	–	–	–	–	–	
e	–	–	4.08	–	–	–	–	–	–	–	–	–	–	
u	–	–	–	5.14	–	–	–	–	–	–	–	–	–	
^*o*	–	–	–	16.67	–	–	–	–	–	–	–	–	–	
^*u*	–	–	–	–	25	–	–	–	–	–	–	–	–	
bh	–	–	–	–	–	8.29	–	–	–	–	–	–	–	–
ch	–	–	–	–	–	–	5.45	9.09	–	–	–	–	–	–
dh	–	–	–	–	–	–	–	10.71	–	–	3.57	–	–	–
jh	–	–	–	–	–	–	33.3	–	–	–	–	–	–	–
kh	–	–	–	–	–	–	–	2.17	4.69	–	–	–	–	–
m	–	–	–	–	–	2.38	–	–	–	5.75	–	–	–	–
p	–	–	–	–	–	9.64	–	3.61	–	–	–	–	–	–
ph	–	–	–	–	–	–	–	–	–	–	20	–	–	–
R	–	–	–	–	–	–	–	–	–	3.57	–	19.64	–	–
T	–	–	–	–	–	–	–	10.21	–	–	–	–	–	–
Th	–	–	–	–	–	–	–	4.35	8.70	–	–	–	8.70	–
th	–	–	–	–	–	–	–	9.68	–	–	–	–	–	–

acoustic model, and it has been used in the subsequent training process. We have used SCTK toolkit[2] for recognition scoring in this experiment.

In Table 1, We have presented phoneme recognition statistics of baseline and AM3 (combination of LDA and MLLT) triphone acoustic model. It is obvious from Table 1 that combination of LDA and MLLT techniques improve recognition accuracy by 1–2% in average. Combination of LDA and MLLT techniques make decoding process faster. To improve the recognition accuracy further, VTLN has been applied with LDA and MLLT in the next section.

5.1.2. Vocal tract length normalization

VTLN will normalize our vocal tract length which also reduce mismatch in acoustic signal. VTLN has been applied alone for speaker normalization is marked as AM6. Moreover, We have applied VTLN with LDA and MLLT in acoustic model AM8 to get further improvement. Warp factor $\alpha > 1$ means compressing the spectrum and $\alpha < 1$ means stretching the spectrum. $\alpha = 1$ stands for no warping. In VTLN, warp factors are selected from discrete values of range 0.8 to 1.2 with step increment 0.05. In Fig. 2, distribution of warp factors of range 0.75–1.2 of young and aged speakers are provided. It has been observed from Fig. 2 that Most of the young speakers are normalized with lower warp factor ($\alpha = 1.2$) whereas most of the aged speakers are also normalized with warp factor $\alpha = 1.2$.

It has been observed in Table 1 that phoneme recognition accuracy is better for acoustic model of aged groups using test data of aged population. We have provided phone recognition accuracy obtained from all acoustic models. As our main aim of this study is to create better recognition for elderly people, we have concentrated more on acoustic model of aged.

In Table 1, phoneme recognition performance achieved from acoustic model AM6 and combined acoustic model AM8 has been provided. Phoneme recognition accuracy improves more by AM8 acoustic model than AM6.

[2] 'Speech recognition scoring toolkit'. http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm.

5.1.3. Maximum likelihood linear regression and maximum a posteriori

In theory, MAP and MLLR are equivalent if each Gaussian is assigned to a single regression class. In practice, it does not happen. There is always an additive effect on ASR performance if those techniques are combined. To do so, first MLLR transformation matrix is derived from the speaker specific adaptation data. Then, transformation matrix is applied to baseline mean. After that model parameters are re-estimated with transformed mean. At last, MAP adaptation is applied to new model parameter set to produce adapted acoustic model parameters. We have applied those two adaptation techniques to the acoustic model, trained on BENG_YO,

Vowels		
Phoneme	AM1	AMLV
A (আ)	87	89.4
ˆA (আঁ)	73.7	84.2
a (অ)	79.5	84.1
E (অ্যা)	72.2	78.9
e (এ)	86.2	90.1
ˆe (এঁ)	80	95.2
i (ই)	91.2	93.2
o (ও)	84.1	88.1
ˆo (ওঁ)	83.3	83.3
oi (ঐ)	92.3	94
ou (ঔ)	90.4	93.2
u (উ)	88	86.9
ˆu (উঁ)	66.7	75

Consonants			
	Phoneme	AM1	AMLV
Labial	b (ব)	65.8	78.4
	bh (ভ)	60.7	65.2
	m (ম)	84.3	83.3
	p (প)	71.8	71.1
	ph (ফ)	57.1	70
Palato-Alveolar	ch (চ)	50.9	60.7
	chh (ছ)	90.9	92.9
	j (জ)	76.2	87.3
	jh (ঝ)	50	50
Retroflex	D (ড)	65.8	66.7
	Dh (ঢ)	70.2	73.1
	T (ট)	73.6	77.3
	Th (ঠ)	60.9	60.9
	R (ড়)	52.8	66.1
Dental and Alveolar	d (দ)	88.1	87.4
	dh (ধ)	61.5	67.9
	t (ত)	79.8	87.5
	th (থ)	84.4	87.1
	n (ন)	90.4	96.1
	r (র)	78.7	86.4
	s (স)	96.5	98.3
	l (ল)	82.3	83
Velar	g (গ)	75.6	86.7
	gh (ঘ)	61.5	58.3
	k (ক)	79.5	88.5
	kh (খ)	78.4	84.8
	Y (য়)	80	81.4
	ˆn (ং)	84	92.3
Glottal	h (হ)	85	84.1

Fig. 3. Phoneme recognition accuracy (%) employing AM1 and AMLV acoustic model and test sample of aged.

Table 4
Word recognition accuracy (standard deviation) of baseline and other normalized and adapted acoustic model with test data of aged people (OT) and young people (YT).

	AM1	AM6	AM2	AMLM	AMLV	AM3	AM7
Young_OT	54.6 (24.1)	56.9 (23.4)	57.1 (24.4)	62.2 (24.7)	63.2 (24.1)	64.2 (24.5)	64.8 (22.6)
Old_OT	70.6 (12.2)	71.8 (12.1)	74.2 (13.4)	77.5 (11.4)	78.1 (11.4)	79.6 (11.4)	81.8 (11.6)
Mixed_OT	60.7 (15.6)	62.3 (15.8)	63.9 (15.2)	70.6 (15)	73.3 (15.1)	74.2 (14.6)	74.8 (14.9)
Young_YT	80.6 (12.1)	81.9 (10.4)	82.3 (10.1)	87.8 (9.8)	88.9 (9.2)	89.2 (8.8)	90.3 (8.5)

BENG_OL and BENG_MIX corpus. MLLR and MAP adapted acoustic model of elderly performs better than baseline system of aged. Here, we have combined VTLN with MLLR and MAP for aiming better accuracy.

We have achieved phoneme accuracy 79.8% in average using combination of [VTLN + MLLR + MAP] whereas baseline system of aged people yield phoneme recognition accuracy 72.6%. Tau is the prior weight, a control parameter in the MAP adaptation. Optimum ASR accuracy has been achieved at tau value near by 12800.

We have also employed VTLN, MLLR and MAP on baseline acoustic model of combined age group. ASR performance of mixed acoustic model adapted with combined MAP and MLLR is experimented. We have tried to fix the value of tau which will provide better ASR accuracy as well as less standard deviation (S.D.).

Adaptation on acoustic model of BENG_YO does not provide satisfactory accuracy for test data of aged population with respect to adapted acoustic model of aged population. We have calculated accuracy of all baseline and adaptive acoustic model. Here, phoneme recognition result achieved by adapted acoustic model of aged population is provided in Table 1.

5.1.4. Maximum mutual information estimation

We have combined LDA, MLLT, VTLN, MLLR, MAP and MMIE techniques as mentioned in Fig. 1 to obtain acoustic model AMLM, AM3, AMLV and AM7. It has been observed from recognition results that MMI based acoustic model performs better for aged population than the other methods. In Table 1, we have provided phoneme recognition accuracy of MMIE model (AMLM) and other speaker adaptive model AMLV, AM3 and AM7. Though MMIE consumes huge amount of time at the time of training, it performs good once acoustic model created.

5.1.5. Overall improvement in accuracy: phone recognition

We will now discuss about the confused phone in baseline and final adaptive system. Confusion matrix of baseline and adaptive system represents amount of confusion by percentage of a phone with other phoneme. Here, most significant phoneme according to confusion will be discussed. From Tables 2 and 3, it can be observed that nasal vowels are confused with its base vowels like $/\hat{}A/ \Rightarrow /A/$, $/\hat{}o/ \Rightarrow /o/$, and $/\hat{}u/ \Rightarrow /u/$. Final speaker adaptive model can reduce the amount of confusion for $/\hat{}A/$ and $/\hat{}u/$ but not for $/\hat{}o/$. Inspite of nasal vowels, /E/ has a tendency to be confused with /A/, and /a/, /u/ are confused with vowel /o/.

There are some consonants also have tendency to be confused with other consonants e.g. /bh/ ⇒ /b/, /ch/ ⇒ /j/, /jh/ ⇒ /j/, /kh/ ⇒ /k/, /ch/ ⇒ /j/, /p/ ⇒ /b/, /T/ ⇒ /t/, and /th/ ⇒ /t/. In Tables 2 and 3, confusion matrix of some consonants are provided.

Phoneme recognition results are organized according to place of articulation. In Fig. 3, recognition pattern of vowels, labial, dental, retrophlex, palato-alveolar, and glottal phoneme are provided for overall realization of phoneme recognition. It can be figured out that labial, retroplex and dental phoneme are affected more than other groups with aging. However, improvement of accuracy has been achieved by combination of LDA, MLLT, VTLN and MMIE based acoustic model.

We can conclude from Fig. 3 that recognition accuracy of vowel /E/, $/\hat{}u/$ and $/\hat{}A/$ are lower than other vowels. Although recognition accuracy improves by exploiting speaker adapted model. However, lower accuracy rate of those three vowels have correlation with aging. Jaw and tongue movement is important to pronounce those vowels. Recognition result shows recognition accuracy of dental phone /dh/ and /t/ are affected more with aging according to recognition accuracy. Although adapted model increase the accuracy of those phoneme. All the labial phoneme shows lower recognition accuracy. However, only /m/ is recognized with better accuracy than other labial phoneme. AMLV acoustic model improves recognition accuracy of /ph/, /b/ and /bh/ by 10–12% in average. Inspite of /gh/, all velar phoneme provide better recognition accuracy (80% in average) as shown in Fig. 3. Recognition performance of retroplex phoneme (e.g. /T/, /Th/ and /D/) are affected more with aging. Although palato-alveolar phoneme (e.g. /chh/, /j/) and glottal phone /h/ are recognized quite well whereas /ch/ and /jh/ are affected by means of accuracy.

5.1.6. Word recognition

Word recognition is the final goal of our study. As word recognition is directly related to phoneme recognition, we have discussed it in earlier section. Words are constructed with affected phoneme yield poor recognition rate. We only describe here the word recognition accuracy of different acoustic model with test data of aged population. However, among several methods has been investigated in this study, adaptation techniques combined with MMIE performs better than other methods. Acoustic model of mixed training data performs better than AM of BENG_YO but lagging faraway from performance of acoustic model of aged.

We have used Mann–Whitney–Wilcoxon (MWW) (Mann and Whitney, 1947) non parametric unpaired test for measuring statistical significance. Improvement of word recognition accuracy obtained from AM7 is statistically significant at P value 0.0001 with respect to AM1. It also has been observed that results obtained from AM2 is statistically significant at P value 0.05.

Table 4 shows, word recognition accuracy exploiting test sample of young age group on different acoustic models of young. We have achieved word recognition accuracy 90.3% on test data of young age group by acoustic model AM7 which is statistically significant with respect to AM1 at P value 0.0001. Although AM7 provides better accuracy, it is a speaker specific model. It only performs better for a particular speaker who has been adapted to the model by means of MLLR and MAP subsequently. Acoustic model AM7 require training samples from respective speaker to adapt the model. It is always an costly task. AMLV can outperform all other acoustic model in speaker independent adaptation methods. In AMLV, test samples are normalized first by VTLN approach, and then transformation matrix obtained from MLLT is applied to Gaussian parameters for speaker adaptation.

6. Conclusion

In this paper, speaker adaptive acoustic models has been developed with the aim of improving speech recognition accuracy of aged population. Recognition experiments has been carried out with two different test set made of same sentences recorded with Adults (20–40 years of age) and aged (60–80 years of age). Trigram language model has also been applied with the acoustic model to incorporate linguistic information.

In the speaker adaptive acoustic modeling, different methods has been experimented such as VTLN, LDA, MLLT, MLLR and MAP. There is a baseline system which is developed with maximum likelihood estimation algorithm. Furthermore, different combination of adaptation techniques e.g. [VTLN + LDA + MLLT], [MAP + MLLR], MMIE and [VTLN + LDA + MLLT + MMIE] are implemented in this study. All these techniques are implemented first on baseline acoustic model of young population and then on aged population. It has been observed that adapted or normalized acoustic model of young group improves the performance with respect to baseline system (AM1) but does not fulfill the desired accuracy. This motivate us to build up the adapted acoustic model of aged group. In every adaptation technique, speech recognition accuracy improves with respect to baseline system. However, best recognition result is obtained with speaker specific adapted acoustic model *AM7*. *AM7* performs well only for specific speaker. We have used *AMLV*, a speaker independent adapted model to overcome specified restriction in *AM7*.

According to phoneme recognition of test sample of aged population, it can be conclude that dental, labial, retroplex consonants are affected much with aging. Phoneme recognition results have shown an important observation about nasal and fricative phoneme. Nasal and fricative phoneme are not affected much with aging.

Acknowledgments

This work is supported by Technology Intervention for Elderly, Department of Science and Technology, Government of India for the project “Elderly speech recognition with applications”.

References

Bahl, L., Brown, P., de Souza, P., Mercer, R., 1986. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. IEEE Internat. Conf. on ICASSP ’86.Acoustics, Speech, and Signal Process. 11, 49–52.

Das, B., Mandal, S., Mitra, P., 2011. Bengali Speech Corpus for Continuous Automatic Speech Recognition System, Internat. Conf. on Speech Database and Assessments (Oriental COCOSDA), pp. 51–55.

Digalakis, V., Rtischev, D., Neumeyer, L., 1995. Speaker adaptation using constrained estimation of gaussian mixtures. IEEE Trans. Speech Audio Process. 3 (5), 357–366.

Eide, E., Gish, H., 1996. A parametric approach to vocal tract length normalization. IEEE Internat, Conf. on Acoustics, Speech, and Signal Process. 1996. ICASSP-96. Conf. Proc. 1, 346–348.

Gales, M., Woodland, P., 1996a. Mean and variance adaptation within the MLLR framework. Comput. Speech Lang. 10, 249–264.

Gales, M.J.F., Woodland, P., 1996b. Variance compensation within MLLR framework. Technical Report. Report CUED/F-INFENG/TR242, Cambridge University, Cambridge, UK, pp. 171–185.

Gauvain, J.-l., Lee, C.-h., 1992. Map Estimation of Continuous Density HMM: Theory and Applications, pp. 185–190.

Giuliani, D., Gerosa, M., Brugnara, F., 2006. Improved automatic speech recognition through speaker normalization. Comput. Speech Lang. 20 (1), 107–123.

Gopinath, R.A., 1998. Maximum Likelihood Modeling with Gaussian Distributions for Classification, pp. 661–664.

Gorham-Rowan, M.M., Laures-Gore, J., 2006. Acoustic-perceptual correlates of voice quality in elderly men and women. J. Comm. Disorders 39 (3), 171–184.

Haeb-Umbach, R., Ney, H., 1992. Linear discriminant analysis for improved large vocabulary continuous speech recognition. IEEE Internat. Conf. on Acoustics, Speech, and Signal Process. 1992. ICASSP-92, 1, 13–16.

Hillenbrand, J., Cleveland, R.A., Erickson, R.L., 1994. Acoustic correlates of breathy vocal quality. J. Speech Lang. Hear. Res. 37 (4), 769–778.

Huang, X., Lee, K., 1991. On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. IEEE Trans. on Speech and Audio Process. 2, 877–880.

Huda, S., Yearwood, J., Togneri, R., 2009a. A constraint-based evolutionary learning approach to the expectation maximization for optimal estimation of the hidden markov model for speech signal modeling. IEEE Trans. on Systems, Man, and Cybernet. Part B: Cybernet. 39 (1), 182–197.

Huda, S., Yearwood, J., Togneri, R., 2009b. A stochastic version of expectation maximization algorithm for better estimation of hidden markov model. Pattern Recognition Lett. 30 (14), 1301–1309.

Krom, G.d., 1993. A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. J. Speech Lang. Hear. Res. 36 (2), 254–266.

Lee, C.-H., Gauvain, J.-L., 1993. Speaker Adaptation Based on Map Estimation of HMM Parameters, IEEE Internat. Conf. on Acoustics, Speech, and Signal Process. 1993. ICASSP-93., pp. 558–561.

Lee, L., Rose, R., 1996. Speaker normalization using efficient frequency warping procedures. IEEE Internat. Conf. on Acoustics, Speech, and Signal Process. 1996. ICASSP-96. Conf. Proc. 1, 353–356.

Lee, K.-F., Hon, H.-W., Hwang, M.-Y., Huang, X., 1990. Speech recognition using hidden markov models: a CMU perspective. Speech Comm. 9 (5G6), 497–508.

Leggetter, C., Woodland, P., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. Comput. Speech Lang. 9 (2), 171–185.

Lindblom, B.E.F., 1971. Acoustical consequences of lip, tongue, jaw, and larynx movement. J. Acoust. Soc. Amer. 50, 1166–1179.

Linville, S.E., 2001. Vocal Aging. Singular Publishing Group, San Diego.

Linville, S.E., Rens, J., 2001. Vocal tract resonance analysis of aging voice using long-term average spectra. J. Voice 15 (3), 323–330.

Liu, F.-h., Stern, R.M., Huang, X., Acero, R., 1993. Efficient Cepstral Normalization for Robust Speech Recognition. In: Proceedings of the workshop on Human Language Technology, pp. 69–74.

Mandal, S., Das, B., Mitra, P., Basu, A., 2011. Developing Bengali speech corpus for phone recognizer using optimum text selection technique. In: Internat. Conf. on Asian Language Processing, pp. 268–271.

Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Statist. 18 (1), 50–60.

Molau, S., Pitz, M., Schluter, R., Ney, H., 2001. Computing mel-frequency cepstral coefficients on the power spectrum. IEEE Internat. Conf.on Acoustics, Speech, and Signal Process. 2001. Procee. (ICASSP '01). 1, 73–76.

Normandin, Y., Lacouture, R., Cardin, R., 1991. MMIE training for large vocabulary continuous speech recognition. ICASSP-91., International Conference on Acoustics, Speech, and Signal Processing, pp.537-540 vol.1, 14–17

Paulsen, F.P., Tillmann, B.N., 1998. Degenerative changes in the human cricoarytenoid joint. Arch. Otolaryngol. Head Neck Surg. 124, 903–906.

Rodeo, M.T., Sánchez-Fernández, J.M., Rivera-Pomar, J.M., 1993. Histochemical and morphometrical ageing changes in human vocal cord muscles. Acta Oto-Laryngol. 113, 445–449.

Rother, P., Wohlgemuth, B., Wolff, W., Rebentrost, I., 2002. Morphometrically observable aging changes in the human tongue. Ann. Anat. – Anat. Anzeiger 184 (2), 159–164.

Tolep, K., Higgins, N., Muza, S., Criner, G., Kelsen, S.G., 1995. Comparison of diaphragm strength between healthy adult elderly and young men. Amer. J. Resp. Crit. Care Med. 152 (2), 677–682.

Torre, P.I., Barlow, J.A., 2009. Age-related changes in acoustic characteristics of adult speech. J. Comm. Disorders 42 (5), 324–333.

Uebel, L.F., Woodland, P.C., 1999. An Investigation into Vocal Tract Length Normalisation. In: Proceedings Eurospeech, vol. 6, pp. 2527–2530.

Ulatowska, H.K., 1985. The Aging Brain: Communication in the Elderly. College-Hill Press, San Diego.

Valtchev, V., Odell, J., Woodland, P., Young, S., 1997. MMIE training of large vocabulary recognition systems. Speech Comm. 22 (4), 303–314.

Vipperla, R., Renals, S., Frankel, J., 2010. Ageing Voices: The Effect of Changes in Voice Parameters on ASR Performance, EURASIP Journal on Audio, Speech, and Music Process. 2010, 5, p. 10.

Wegmann, S., McAllaster, D., Orloff, J., Peskin, B., 1996. Speaker normalization on conversational telephone speech. Proceedings of the Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings., 1996 IEEE International Conference.1, 339–341.

Welling, L., Kanthak, S., Ney, H., 1999. Improved methods for vocal tract normalization. IEEE Internat. Conf. on Acoustics, Speech, and Signal Process. 1999. Proc. 2, 761–764.

Wilcox, K.A., Horii, Y., 1980. Age and changes in vocal jitter. J. Gerontol. 35 (2), 194–198.

Wilpon, J., Jacobsen, C., 1996. A study of speech recognition for children and the elderly. IEEE Internat. Conf. on. Acoustics, Speech, and Signal Process. 1996. ICASSP-96. Conf.Proc. 1, 349–352.

Xue, S.A., Hao, G.J., 2003. Changes in the human vocal tract due to aging and the acoustic correlates of speech production: a pilot study. J. Speech Lang. Hear. Res. 46 (3), 689–701.

Yumoto, E., Sasaki, Y., Okamura, H., 1984. Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness. J. Speech Lang. Hear. Res. 27 (1), 2–6.