Towards Enhancing the Acoustic Models for Dysarthric Speech

Kuruvachan K. George and C. Santhosh Kumar

Machine Intelligence Research Lab., Amrita School of Engineering, Amritanagar, Coimbatore, India
kg_kuruvachan@cb.amrita.edu

Abstract. Dysarthria is a set of congenital and traumatic neuromotor disorders that impair the physical production of speech. These impairments reduce or remove the normal control of the vocal articulators. The acoustic characteristics of dysarthric speech is very different from the speech signal collected from a normative population, with relatively larger intra-speaker inconsistencies in the temporal dynamics of the dysarthric speech [1] [2]. These inconsistencies result in poor audible quality for the dysarthric speech, and in low phone/speech recognition accuracy. Further, collecting and labeling the dysarthric speech is extremely difficult considering the small number of people with these disorders, and the difficulty in labeling the database due to the poor quality of the speech. Hence, it would be of great interest to explore on how to improve the efficiency of the acoustic models built on small dysarthric speech databases such as Nemours [3], or use speech databases collected from a normative population to build acoustic models for dysarthric speakers. In this work, we explore the latter approach.

1 Introduction

Dysarthria [4] is a speech disorder due to a brain, nerve or muscle damage resulting in lack of control on the muscles of tongue, mouth, larynx or vocal cords that produce speech. The muscles may be weak, completely paralyzed, or the coordination between them might have failed. The speech of dysarthric patients is poorly audible, improperly pronounced, or without any rhythm or speed and of very poor quality. Due to the poor quality of dysarthric speech data, the performance of a speech recognition system build on speech data collected from the normative population will be very bad.

In most kinds of motor speech disorders articulatory gestures are typically slow, even when the speaking syllable rate is faster than normal as in the case of dysarthria associated with Parkinsons disease [1]. Speech temporal impairments can include unclear distinction between adjacent phonemes due to imprecise placement of articulators, slower speech rates, and rhythmic disturbances, to name a few [2]. Thus, it may be seen that the distorted temporal dynamics of the speech signal is one of the important reasons that causes degradation in the quality of dysarthric speech[2]. Further, it is also observed that formant trajectories of dysarthric patients are inconsistent across repetitions [1].

V.G. Duffy (Ed.): DHM/HCII 2013, Part I, LNCS 8025, pp. 183–188, 2013.

There are many difficulties in building a good quality acoustic model for dysarthric speech:

1. Collecting large dysarthric database is extremely difficult, due to the small percentage of the population with this disorder.
2. Labeling dysarthric speech is extremely difficult due to the poor quality of the speech signal.

Yet, the importance of acoustic models for computer assisted recognition of dysarthric speech is no less important. Usually, people with dysarthria needs treatment from specialist clinics, mainly for the therapy sessions to improve the patient's speaking skills. Often, this means traveling long distances for the therapy. A good quality acoustic model, and fairly reasonable recognition accuracy can help build low cost computer assisted therapy tools for the dysarthric patients.

Since it is difficult to get large dysarthric speech databases for languages across the world, it would be of great interest if we can build acoustic models using speech databases of the normative population, and then transform these models to dysarthric speech for improved performance.

2 Maximum A Posteriori (MAP) Adaptation

Maximum a posteriori adaptation(MAP)[5] can be used for adapting the existing phone models. MAP is a speaker independent adaptation technique. It uses prior knowledge about the model parameters to adapt the present models. In this approach, model parameters are estimated and modified in such a way that the likelihood of the adaptation data to be generated by the adapted model is maximized. In MAP, the model parameter estimate is considered as a random variable, which has prior probability distribution. Using this prior probability, we calculate the posterior probability, the maximum posterior probability is considered as the adapted model estimates. The new model with adapted model parameters are generated with the MAP estimate, utilizing the knowledge about the prior parameters, weights and the adaptation data.

The update formula for a single stream system for state j and mixture component m is,

$$\hat{\mu}_{jm} = \frac{N_{jm}}{N_{jm} + \tau}\bar{\mu}_{jm} + \frac{\tau}{N_{jm} + \tau}\mu_{jm} \tag{1}$$

where τ is a weighting of the a priori knowledge to the adaptation speech data and N is the occupation likelihood of the adaptation data, defined as,

$$N_{jm} = \sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t) \tag{2}$$

where μ_{jm} is the speaker independent mean and $\bar{\mu}_{jm}$ is the mean of the observed adaptation data and is defined as,

$$\bar{\mu}_{jm} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t) o_t^r}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t)} \tag{3}$$

The amount of adaptation data required for MAP adaptation is relatively high. The performance goes down as the amount of adaptation data available becomes less. In this work the MAP adaptation is performed for the phone models trained on the speech data collected form normative population feeding sufficient amount of the adaptation data collected from dysarthric patients.

3 Maximum Likelihood Linear Regression(MLLR)

Unlike MAP adaptation, Maximum Likelihood Linear Regression(MLLR)[6] needs relatively less amount of adaptation data, and is usually used for speaker adaptation of the speaker independent acoustic model. In this technique, the parameters of the speaker independent (MAP adapted) phone models are modified based on the linear transformation (regression) matrix . MLLR adaptation makes use of the available adaptation data to formulate the regression matrix. This regression matrix is calculated and refined using the forward-backward algorithm [7] and it maximizes the likelihood of the adaptation data.

A particular distribution, s, is characterized by a mean vector, μ_s and a covariance matrix C_s. Given a parameterized speech frame vector o, the probability density of that vector being generated by distribution s is $b_s(o)$

$$b_s(o) = \frac{1}{(2\pi)^{n/2}|C_s|^{1/2}} e^{1/2(o-\mu_s)'C_s^{-1}(o-\mu_s)} \tag{4}$$

where n is the dimension of the observation vector.

The adaptation of the mean vector is achieved by applying a regression matrix W_s to the extended mean vector ξ_s to obtain an adapted mean vector $\hat{\mu_s}$

$$\hat{\mu_s} = W_s \xi_s \tag{5}$$

where W_s is an $n \times (n+1)$ matrix which maximizes the likelihood of the adaptation data, and ξ_s is defined as

$$\xi_s = [\omega, \mu_1, ..., \mu_n]' \tag{6}$$

where ω is the offset term for the regression.

For distribution s, the probability density function for the adapted system becomes

$$b_s(o) = \frac{1}{(2\pi)^{n/2}|C_s|^{1/2}} e^{1/2(o-W_s\xi_s)'C_s^{-1}(o-W_s\xi_s)} \tag{7}$$

In our work, we use MLLR technique to adapt only the mean vectors of the MAP adapted phone models as it requires relatively less training data compared to adapting mean, variance and mixture weights.

4 Experiments and Results

Our baseline system uses mel frequency cepstral coefficients(MFCC), with zero mean subtraction, and delta and acceleration coefficients appended. We use 13

MFCC coefficients, and this makes the total number of features in the acoustic model to 39. For training the acoustic models, we split the data speakerwise into training and test set. Training data consisted of data from speakers, BB, BK, BV, FB, JF, KS, LL. Speakers MH, RK, RL, and SC are used for testing. 20 per cent of the speech from every speaker is used for adaptation of the acoustic models, and the remaining 80 per cent for testing. Nemours database is phonetically transcribed using the TIMIT[1] transcriptions. The database is recorded at 16 kHz. Acoustic models trained on the training data are MLLR adapted using the adaptation data to generate the speaker dependent acoustic models. For small amounts of adaptation data, MAP adaptation was seen to be ineffective, and hence not MAP adaptation was not performed on this acoustic model. Table 1 lists the phone recognition accuracy of the acoustic model trained on Nemours data and tested on Nemours data, with and without adaptation.

Table 1. Phone recognition accuracy of the baseline acoustic model trained using the TIMIT, and Nemours databases on the Nemours test set

Train data	Adaptation	Accuracy
NEMOURS	NO	31.83%
NEMOURS	MLLR	36.31%

Next, we built another acoustic model using the TIMIT database, recorded at 16 kHz desktop quality speech. We then MAP adapted this acoustic model using the Nemours training data to transform the acoustic model from the normative population feature space to the dysarthric feature space. It was observed that the performance of the MAP adapted acoustic model is better than the baseline acoustic models trained using the Nemours database. Fig.1 illustrates schematically how different adaptation techniques are applied on the baseline system built using the TIMIT database. Table 2 lists the performance of the acoustic model trained on TIMIT, when tested without any adaptation, and with MAP, and MAP+MLLR adaptation using the Nemours data.

Table 2. Phone recognition accuracy after MLLR adaptation

Train data	Adaptation	Accuracy
TIMIT	NO	29.69%
TIMIT	MAP	32.67%
TIMIT	MAP + MLLR	39.61%

It may be noted from Tables 1 and 2 that training the acoustic model on speech data from a normative population, and adapting the models to the dysarthric

[1] `http://www.ldc.upenn.edu/Catalog/LDC93S1.html`

speech is better than training the models on the dysarthric speech. This is perhaps due to the ability of the acoustic models built using TIMIT to learn the consistent aspects of the speech better than the acoustic model trained using the Nemours database. Subsequently, adaptation transforms the models to the Nemours feature space for improved performance.

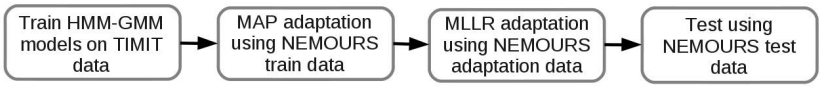

Fig. 1. Sequence of model training, adaptations and testing

All phone recognition experiments in this work are performed using a phone loop without using any language models. All phones has three states each, and evry phone state use 64 Gaussians to model the probability distribution, as this was found to be the optimum configuration empirically. No Gaussians are shared between phone states [8][9].

It may be noted that the performance of the acoustic models may be further enhanced by using bigram/trigram phone language models, or performing the experiments for word recognition with word language models. Using triphone acoustic models also may considerably enhance the phone recognition accuracy of the acoustic models.

5 Conclusion

Dysarthric phone/speech recognition has always been very challenging mainly due to the unavailability of enough amount of well labeled speech data, and the poor quality of the dysarthric speech. Often getting a labelled dyarthric speech database for many of the world languages is extremely difficult, if not impossible. It would be of great interest to explore building acoustic models using speech from a normative population that is much easily available, and then adapt these acoustic models using a small amount of dysarthric speech. It was seen that for small amounts of speech data, it is advantageous to train acoustic models on speech from a normative population and then adapt the acoustic models to the dysarthric feature space. By doing this, we may be able to get an improved phone recognition accuracy over a phone recognition system built on the dysarthric speech alone.

References

1. Weismer, G., Tjaden, K., Kent, R.D.: Can articulatory behavior in motor speech disorders be accounted for by theories of normal speech production? Journal of Phonetics 23, 149–164 (1995)

2. Duffy, J.: Motor Speech Disorders: Substrates, Differential Diagnosis, and Management. Mosby, St. Louis (2005)
3. Menendez-Pidal, X., Polikoff, J.B., Peters, S.M., Leonzio, J.E., Bunnell, H.T.: The Nemours database of Dysarthric speech. In: Proceedings of the Fourth International Conference on Spoken Language Processing, Philadelphia, USA (1996)
4. Murdoch, B.E. (ed.): Dysarthria: A Physiological Approach to Assessment and Treatment, ch. 1. Stanley Thornes Publishers Ltd., UK (1998)
5. Gauvain, J.L., Lee, C.H.: Maximum a posteriori estimation for multivariate Gaussian mixture obser-vations of Markov chains. IEEE Transactions on Speech and Audio Processing 2, 291–298 (1994)
6. Leggetter, C.J., Woodland, P.C.: Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language 9, 171–185 (1994)
7. Young, S., Jansen, J., Odell, J., Ollason, D., Woodland, P.: The HTK book. Cambridge University Engineering Department, Cambridge (2003)
8. Deller, J.R., Hsu, D., Ferrier, L.J.: On the use of Hidden Markov Modelling for recognition of dysarthric speech. Computer Methods and Programs in Biomedicine 35, 125–139 (1991)
9. Reynolds, D.A.: A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification. Ph.D. thesis, Georgia Institute of Technology (1992)