AUTOMATIC CLASSIFICATION OF VOICE QUALITY: COMPARING REGRESSION MODELS AND HIDDEN MARKOV MODELS

Mirjam Wester
Department of Language & Speech, University of Nijmegen
P.O. Box 9103, 6500HD Nijmegen, the Netherlands
wester@let.kun.nl

ABSTRACT

In this paper, two methods for automatically classifying voice quality are compared: regression analysis and hidden Markov models (HMMs). The findings of this research show that HMMs can be used to classify voice quality. The HMMs performed better than the regression models in classifying breathiness and overall degree of deviance, and the two methods showed similar results on the roughness scale. However, the results are not spectacular. This is mainly due to the type of material that was available and the number of listeners who assessed the material. Nonetheless, I argue in this paper that these findings are interesting because they are a promising step towards developing a system for classifying voice quality.

1. INTRODUCTION

Automatically evaluating voice quality could be useful in a clinical practice. It is known that listeners are inconsistent in judging pathological voice quality. One listener can give a different assessment of the same speech fragment on two separate occasions. The judgements of two expert listeners can also be different. Due to these inconsistencies, the perceptive evaluation of voice quality is not very reliable [1,2]. If the knowledge of a number of experienced listeners is stored in such a way that a speech therapist could have access to it through a machine, it will be possible to achieve more consistent and reliable judgements of voice quality.

Listeners base their voice quality judgements on acoustic information in the signal. We can therefore assume there are systematic acoustic differences between voices which make one voice sound healthy and another extremely breathy, for instance. By combining the acoustic information in a speech signal with information about the perceptual evaluation of voice quality, a set of models can be made, which can be used to classify voice quality.

This research is limited to determining if models based on acoustic information in the signal and perceptive labels given by listeners are suitable for classifying voice quality. In order to do so, I compared the automatic classification of voice quality using regression models to the automatic classification of voice quality using HMMs.

The material which was used is described in section 2, followed by a detailed explanation of the two methods. Subsequently, the results obtained with both methods are discussed in section 3. Conclusions are given in section 4.

2. METHOD AND MATERIAL

2.1 Material

The speech material used in these series of experiments consists of 643 fragments; 607 fragments were recordings of pathological voices and 36 of normal voices. The recordings were extracted from the Kay Elemetrics CD-ROM [3]. Two kinds of speech material were available for each speaker: sustained vowels and read speech. The read speech consisted of the first 12 seconds of the “Rainbow Passage”.

It has been shown [1,4] that dynamic variation in the speech signal is especially important for the perception of voice quality. For this reason, the sustained vowels were not suitable. Only the steady-state part of the vowel had been put on the CD-ROM. The onset and offset were missing. Alternatively, the phoneme sequence /ʌnlai/ was segmented from the word ‘sunlight’, which was extracted from the read speech. I chose this fragment because there is a transition from the voiceless /s/ to the voiced fragment /ʌnlai/, which is then followed by the unvoiced /t/. As a result, the relevant transient phenomena are present in this fragment, which makes it suitable for the present experiments.

2.1.1 Perceptive labels

All of the material was assessed by three experienced listeners. The listeners were asked to

label each fragment on three different scales: roughness, breathiness, and general degree of deviance. The scales were all 5-point scales, ranging from non-deviant to extremely deviant. Thus, each fragment was assigned three labels by each listener.

An average of the listener's scores was needed in order to get one perceptive label for each fragment on each scale. A regression analysis was performed to measure whether the listeners had labelled the material consistently on the basis of the acoustic information in the signal. For one of the listeners, the percentage of explained variance was 10% lower than for the other two listeners, on all three scales. Therefore, the scores given by this listener were not included in labelling the speech fragments. Next, a two-sample analysis was performed on the perceptive labels of the two remaining listeners in order to find the difference in means. The difference in means has been incorporated in the average perceptive label for each fragment. Combining the judgements of the two listeners leads to a greater percentage of variance explained than the separate labels do.

The percentage of variance explained for the combined perceptive labels of the two listeners is shown in Table 1. The predictors resulting from the regression analysis are also shown, together with the equations for the average perceptive labels for listener one and two. In the next section, the acoustic parameters are explained.

Table 1: Results of the regression analysis on total corpus for listener 1 and 2 combined for each voice quality scale
%ve = percentage of variance explained

scale	%ve	predictors, acoustic parameters
general deviance	50	HNR1, HNR2, HNR3, low, mid, high slope, lnF0 (gen1+(gen2 +0.32))/2
breathiness	63	HNR1, HNR3, HNR4, mid, high slope, lnF0 (breathy1+(breathy2 + 0.176))/2
roughness	40	HNR1, HNR2, mid slope, lnF0, leveldB ((rough1 + 0.6) + rough2)/2

2.1.2 Acoustic Analysis

An acoustic analysis was carried out [5], in which measurements were made of :

- the Harmonics-to-Noise Ratio (HNR) in four non-overlapping frequency bands,
  HNR1: 60-400Hz,
  HNR2: 400-2000Hz,
  HNR3: 2000-5000Hz,
  HNR4: 5000-8000Hz,
- the differences in level between these frequency bands as an indicator of spectral slope,
  HNR4-HNR3 = high slope
  HNR3-HNR2 = mid slope
  HNR2-HNR1 = low slope.
- the fundamental frequency (lnF0),
- the overall intensity of the signal(leveldB).

The parameters were estimated frame by frame, every 10ms.

2.1.3 Training and test material

Distribution of the material in training and test corpora was done on the basis of the perceptive labels given by the listeners. Each voice quality scale was divided into five parts. Each fragment was assigned a perceptive label between 0 and 4 based on the labels given by the two listeners. For each voice quality scale, 80% of the material was used as training material and 20% was used for testing.

2.2 Method

In this section, I will first explain how the regression models were obtained. Then, I will explain how the HMMs were designed and selected for testing. Finally, I will explain why, in theory, HMMs should perform better in classifying voice quality than regression models.

2.2.1 Regression models

Regression analysis has traditionally been used as a method for the objective evaluation of voice quality [6]. In regression analysis, the relationship between a dependent variable and one or more independent variables is studied. When this method is used for analysing voice quality, the independent variables are acoustic parameters which are used to predict the dependent variables, voice quality labels, given by human experts.

The formula for a linear regression model is:

$$Label = c_0 + c_1 \overline{P_1} + c_2 \overline{P_2} .... + c_n \overline{P_n}$$

The values of the regression coefficients $c_1 ... c_n$ are estimated on the basis of the training material. P_n is the average value of an acoustic parameter over time. The correlation between the perceptive labels given by the listeners and the predicted labels resulting from the regression model can be used as

an indicator of the validity of the model.

In the experiments reported here, “stepwise regression” (Statistical Graphics Corporation) was used to perform the regression analysis. “Stepwise regression” is an automatic way to decide which independent variables should be included in the regression model and which variables are redundant and can consequently be discarded. The acoustic parameters are: HNR1 - HNR4, low, mid, and high slope, lnF0, and leveldB. The perceptive labels are the dependent variables.

Table 2 shows the predictors and corresponding regression coefficients which resulted from the “stepwise regression” for each voice quality scale. A ‘-’ means the predictor is redundant for the corresponding voice quality scale.

Table 2: Predictors and regression coefficients for three voice quality scales.

voice quality scale	general deviance	breathiness	roughness
Predictors	Regression coefficients		
Constant (c_0)	-1,67	-2,89	4,48
HNR1	-0,07	-0,03	-0,04
HNR2	-	-	-0,04
HNR3	-0,14	-0,18	-
HNR4	-	-0,18	-
mid slope	-0,04	-0,09	0,04
high slope	-0,03	-0,06	-
lnF0	3,02	3,51	-0,85
leveldB	-	-	0,02

The signs of the regression coefficients in Table 2 are in line with findings reported in [1]. The influence of the signs may be clarified by the following examples. The lnF0 regression coefficient is negative for roughness but positive for general degree of deviance and breathiness. This means that an increase in lnF0 leads to a lower predicted label for roughness, but to a higher predicted label for breathiness and general degree of deviance. A higher predicted label means the fragment is more rough, breathy, or deviant. Another example is the negative regression coefficients for the HNRs on all voice quality scales. This means that a high HNR leads to a low voice quality label on all three scales. A high HNR also means that many harmonics and little noise are present in the signal, which is an indication of a healthy voice.

2.2.2 HMMs

A possible alternative to regression models is HMMs. HMMs are probabilistic models which can be used to model a series of observations. Nowadays, they are widely employed in automatic speech recognition [7]. The units of recognition are usually phonemes or words. In this research, the listener's perceptive labels were used as the unit of recognition. This task is comparable to whole word recognition. Essentially, a series of acoustic observations must be associated with the correct HMM. If the HMM matches the correct label, recognition has been successful.

First, a number of prototype HMMs needed to be created. This was done using the Hidden Markov Toolkit V1.5 [8]. For each voice quality scale prototype, HMMs were designed in order to test which parameter settings were optimal for each type of scale. The parameters within a prototype are the number of states, the number of possible skips, and the number of data streams.

The number of states referred to in this paper are the effective states. The entry and exit states in HTK are non-emitting. Therefore, they are not included in the number of effective states. Skips refer to the possible transitions. In a prototype which allows for one skip, self transition and transition to the state on the right is possible. Two skips means that the following state may also be skipped. Multiple data streams are used to enable separate modelling of multiple information sources.

Twenty-one prototype HMMs were designed for each voice quality scale. The number of effective states ranged from one to eight states, with one, two, or three possible skips. Four data streams were defined for all prototype HMMs. They are: one for the HNRs in the four frequency bands, a stream for the low, mid, and high slopes, a stream for lnF0, and a stream for the overall intensity.

Next, the prototype HMMs were trained using the Baum-Welch algorithm [5]. After training, recognition was carried out. During a recognition, every fragment is compared to the five possible models of a voice quality scale. The most probable label is recognized and compared to the perceptive label. If the recognized label matches the perceptive label, recognition is successful.

Two criteria were used to decide which prototype HMM was “the best” for each voice quality scale. First of all, the percentage of correctly classified fragments was taken into account; the second criterion was that if there was more than one model with similar scores, the simplest model was chosen. As a result, the HMM which scored highest

and is still relatively simple was chosen for each voice quality scale.

On the basis of these criteria, the following HMMs were selected. An HMM with four effective states, allowing for two skips, was the best HMM for breathiness. The prototype which scored best for roughness was an HMM with five states and two skips. For the general degree of deviance scale, the HMM chosen consisted of six effective states, allowing for three skips. These three prototype models were used during the rest of the experiments in comparing regression models to HMMs.

2.2.3 HMMs versus regression models

In theory, there are three reasons why HMMs should be able to model voice quality more precisely than regression models. First of all, HMMs are better suited for modelling dynamic variation over time than regression models. Earlier mentioned onset and offset phenomena can be modelled better in an HMM because it consists of a number of states. A second advantage of HMMs opposed to the more traditional regression models is that the input for the HMMs is calculated every 10 ms, whereas the regression analysis calculates one average value over the whole time frame for each parameter. The third reason HMMs are more suitable for classifying voice quality than regression models is because a separate HMM is defined for each point on the voice quality scale, whereas in a regression model the relations are defined over the whole range of a voice quality scale. It is difficult to model non-linear relations over the whole range of a scale using numerous predictors. This problem can be solved by using HMMs because only a part of the range is modelled by each HMM. On the basis of these arguments, HMMs are expected to outperform regression models in classifying voice quality.

3. RESULTS

The previous section explained how the regression models and HMMs were designed. This section presents the results of evaluating the two methods.

The HMMs were evaluated by performing a recognition task. Each fragment was classified by the HMMs as being one of the five categories on the three voice quality scales: breathiness, roughness, and general degree of deviance. The regression models generated predicted labels for each fragment, which can also be viewed as a classification task. The correspondence between classified labels and the perceptive labels given by the listeners leads to a percentage of correctly classified fragments. Besides percentage of correctly classified fragments the percentage of variance explained was also calculated.

There were six different corpora: a training corpus and a test corpus for the three voice quality scales. Figure 1 shows the percentage of correctly classified fragments on the training data. On all three scales the HMMs performed better than the regression models (general deviance +14, breathiness +7%, and roughness +12%).

Figure 2 shows the percentage of variance explained by the two methods on the three voice quality scales measured on the training material. On the general deviance scale (+8%) and the roughness scale (+6%) a higher percentage of explained variance was obtained for the HMMs than for the regression models. On the breathiness scale, the variance explained is a few percent lower (-4%) for the HMMs than for the regression models .

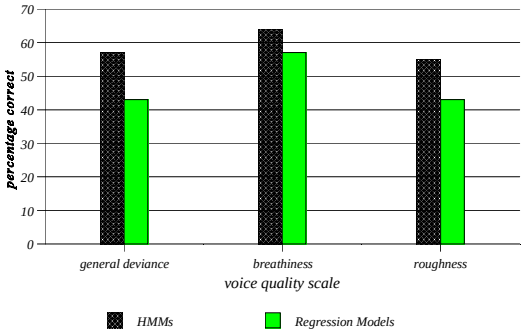

Figure 1: Percentage of correctly classified fragments by the HMMs and the regression models on three voice quality scales (training corpus).

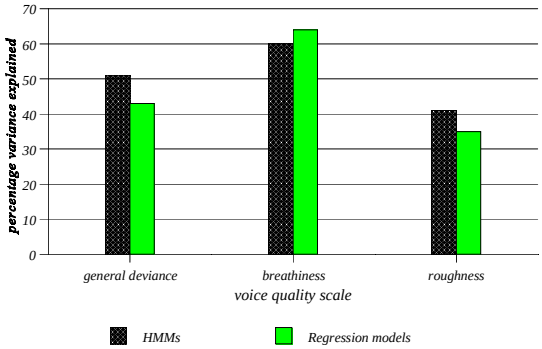

Figure 2: Percentage of variance explained by the HMMs and regression models for each voice quality scale (training corpus).

Figures 3 and 4 show the results obtained on the testing material. The HMMs for general deviance show a higher percentage of correctly classified fragments (+10%) and a higher percentage of variance explained (+12%) than the regression models. The HMMs for breathiness have a higher percentage of correctly classified fragments, but the explained variance is somewhat lower. Finally, the HMMs perform worse than the regression models on the roughness scale on the same data.

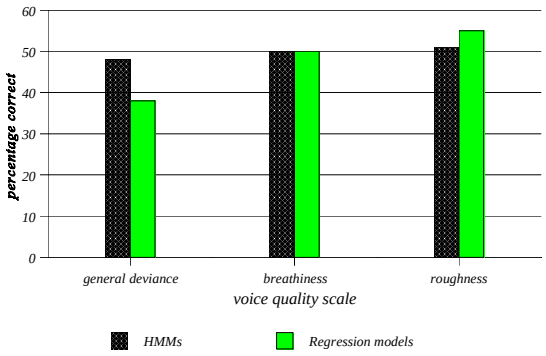

Figure 3: Percentage of correctly classified fragments by the HMMs and the regression models on three voice quality scales (test corpus).

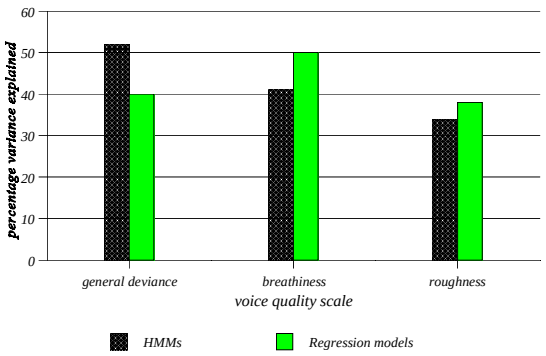

Figure 4: Percentage of variance explained by the HMMs and regression models for each voice quality scale (test corpus).

4. DISCUSSION AND CONCLUSIONS

The findings of this research show that HMMs can be used to classify voice quality. The HMMs performed better than the regression models in classifying breathiness and overall degree of deviance, and the two methods showed similar results on the roughness scale. However, the results are not spectacular. This is mainly due to the type of material which was available and the perceptive labels given by listeners.

For a number of categories there were insufficient fragments available, in order to train the HMMs properly. For example, the category “rough 0" consisted of 262 fragments which were used to train the “HMM-rough 0", whereas there were only 11 fragments for “rough 4". It is obvious that the model for “rough 4" will be undertrained. It would have been better if there had been a more homogeneous division of the material, so that every category could have been trained with the same number of fragments. However, this was not possible because of a lack of available material.

Another main problem in this research is the manner in which the perceptive labels have been assigned to the fragments. In this research, the perceptive labels were chosen on the basis of judgements of voice quality made by only two listeners.

To get a clearer picture of how well the fragments were actually assessed by the listeners, the correlation between the two listeners for all of the material was calculated. Additionally, the percentage of equal labels assigned by the two listeners was calculated. In 40% - 45% of the material, the two listeners assigned the same label to a fragment. The percentage of shared variance between the listeners was 55% on the general degree of deviance scale. This is also the scale on which HMMs score better than the regression models in all cases. For the breathiness, and roughness scales, the shared variance is lower, 47% and 27%, respectively. On these two scales, HMMs perform somewhat worse than the regression models. On the basis of this it can be concluded that the perceptive labels for roughness and breathiness are less accurate than the perceptive labels for general degree of deviance.

If the perceptive labels are more accurate, this means there is more similarity between the labels and the acoustic information in the signal. Models which are trained using more accurate labels will also be able to classify more precisely. In order to obtain correct labelling of speech fragments, a number of expert listeners, preferably more than three, should listen to the material a number of times on different occasions. Only then can the problem of inconsequent labelling be overcome.

In conclusion, a lot of work still needs to be done before a system can be designed which can evaluate voice quality for use in a clinical situation. Whether HMMs are truly suitable for use in such a system

needs to be investigated to a greater extent. Better material and more accurate perceptive labels are prerequisites to designing such a system.

5. REFERENCES

[1] de Krom, G. (1994). Consistency and reliability of voice quality ratings for different types of speech fragments. *Journal of Speech and Hearing Research*, **37**, 985-1000.

[2] Gerratt, B.R., Kreiman, J., Antoñanzas-Barroso, N., & Berke, G.S. (1993). Comparing internal and external standards in voice quality judgments. *Journal of Speech and Hearing Research*, **36**, 14-20.

[3] Kay Elemetrics Corporation (1995). *The Disordered Voice Database version 1.03*.

[4] de Krom, G. (1995). Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *Journal of Speech and Hearing Research*, **38**, 794-811.

[5] de Krom, G. (1993). A cepstrum based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech and Hearing Research*, **36**, 254-266.

[6] Arends, N., Povel,D.J., van Os, E., & Speth, L. (1990). Predicting voice quality of deaf speakers on the basis of glottal characteristics. *Journal of Speech and Hearing Research*, **33**, 116-122.

[7]Huang, X.D., Ariki, Y., & Jack, M.A. (1990). *Hidden Markov Models for Speech Recognition*, Edinburgh: Edinburgh University Press.

[8] Young, S.J., Woodland, P.C., & Byrne, W.J. (1993). *HTK: Hidden Markov Toolkit V1.5*. Entropic Research Laboratories Inc., 600 Pennsylvania Ave. SE, suite 202, Washington, DC 20003 USA.