Extracting voice quality contours using discrete hidden Markov models

Marko Lugger, Frank Stimm, and Bin Yang

Chair of system theory and signal processing
University of Stuttgart, Germany
{marko.lugger; bin.yang}@lss.uni-stuttgart.de

Abstract

In this paper we present an approach of extracting voice quality contours from speech utterances. We apply the theory of hidden Markov models to voice quality classification. As in the case of automatic speech recognition, where the states of the model are interpreted as different phonemes, we interpret the states of our voice quality models as different phonation types. Since non-modal voice quality is only selectively applied in natural speech, the task is to detect those regions within an utterance where these voice qualities were used by the human speech production. We realize that by building so called voice quality contours. Each segment of speech is associated by one discrete voice quality class defined by J. Laver [1]. In this study we distinguish between modal, breathy, creaky, and rough voice.

1. Introduction

Voice quality is an often used instrument to express linguistic and paralinguistic properties in speech. Besides intonation, intensity, and duration, voice quality is one of the important factors of prosody. Former investigations have shown that the voice quality parameters allow the distinction between different speaker groups or speaking styles like gender [2], pathological and non-pathological speakers [3], and word stress [4]. Also, for emotion recognition, a lot of information about the emotional state of a speaker is coded in voice quality aspects [5]. Other potential applications in speech analysis are aging group detection, forensic speaker identification [6], and the improvement of automatic speech recognition (ASR).

Thus an automatic detection of different voice qualities from the speech would be desirable. Our former studies showed the potential of voice quality parameters for the classification of whole utterances [7]. But a segmental classification in terms of a voice quality contour would be more appropriate for characterizing the voice quality content of a spoken utterance. Short parts of an utterance containing only a few segments that are produced in creaky, breathy or rough voice should also be detected by such a system.

There are various methods in the literature for the classification of time series. The most popular method is hidden Markov modelling because of its usage in ASR. We apply discrete hidden Markov models (HMM) for the classification of voice quality. By associating every speech segment with the state of the most likely state sequence of the model, we extract a discrete voice quality contour for a spoken utterance.

The paper is structured as follows. Section 2 defines voice quality and describes the voice quality parameters used in this paper. The well known k-means algorithm for clustering is introduced in section 3. The theory of HMM and its application to the extraction of voice quality contours are presented in section 4. Section 5 shows some results for classifying voice qualities by using HMM. The paper ends up with a conclusion.

2. Voice quality

Voice quality is mainly affected by the excitation of the human voice that is called phonation. Thus, the shape of the glottal pulse is responsible for the voice quality that a speaker is realizing. In contrast, all procedures that belong to the articulation process affect the generated sounds, which all together build the linguistic content.

Usually voice quality parameters are obtained from the electroglottographical signal which is measured at the glottis. Electroglottography (EGG) is a technique used to record the laryngeal behaviour indirectly by measuring the change in electrical conductivity across the throat. There are other methods like stroboscopy or laryngoscopy that belong to medical imaging.

2.1. Voice quality parameters

We use only the acoustic speech signal for voice quality extraction. From the literature, we know a large number of methods for parameterizing voice quality. Time domain parameters like open quotient or skewness quotient are directly related to the parameters from EGG. Another group of methods describe voice quality in the frequency domain mainly by measuring different parameters of spectral tilt.

We follow an approach in the frequency domain by estimating spectral gradients from the glottal excitation spectrum. Therefore, we first compensate the influence of the vocal tract by inverse filtering. It is based on the method first introduced by Stevens and Hanson [8] later extended in [5]. The following 4 spectral gradients are illustrated in Fig. 1 with respect to the pitch frequency F_0: "**O**pen **Q**uotient **G**radient", "**G**lottal **O**pening **G**radient", "**SK**ewness **G**radient", and "**R**ate of **C**losure **G**radient".

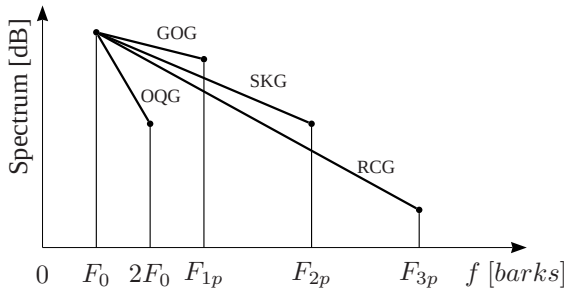

Figure 1: Spectral gradients

F_{kp} are the nearest harmonics to the first 3 formants. There are two more parameters used as features here. These are the pitch frequency F_0 and the bandwidth of the first formant normalized to its frequency, called "**I**ncompleteness of **C**losure". Thus, for every analysis segment of length 25 ms, where periodic excitation is observed, a 6-dimensional voice quality parameter vector is extracted. The overlapping segments are calculated every 10 ms. In the following, solely this parameter vector is used as feature vector $\underline{x}(t)$ for classification.

2.2. Phonation types by J. Laver

John Laver defined a set of discrete voice qualities describing the most important phonation types used in human speech production [1]. These are modal, breathy, creaky, rough, whispery, and falsetto voice. Since the latter two are really sparse in regular language use, only the first four are studied in this paper.

3. Vector quantization

The one-dimensional quantization of scalar values can be extended to k-dimensional vectors. The commonly used method is the k-means clustering algorithm. Vector quantization can also be seen as a classification process which assigns a one-dimensional symbol to a k-dimensional vector. In this paper we use the k-means algorithm for two purposes. On the one hand we use the k-means algorithm to quantize the feature vectors $\underline{x}(t)$ to the discrete observations $o(t)$ which are required as input for the discrete hidden Markov model, see section 4. On the other hand it is directly used for the classification of four voice qualities, see section 5.1. Therefore all the feature vectors are partitioned to 4 clusters. Each cluster is directly associated with a voice quality.

4. Hidden Markov models

Hidden Markov models are well known from [9]. In combination with mel frequency cepstral coefficients (MFCC) they build the state of the art technique in ASR. But HMM can also be used for modelling or classification in other applications. We apply them to voice quality classification.

4.1. Definition of HMM

An HMM is a finite state machine which consists of N states. For every observation $o(t)$ of an observation sequence **O** an underlying state $q(t)$ is assumed. In each state symbols with a corresponding probability are emitted. So, $b_j(k)$ is the probability for being in state j and emitting the symbol k. For every time instant the state of the HMM can change. The probability for a transition from state i to state j is a_{ij}.

4.2. Classifying with HMM

In general classification applications, for each class an own model is generated during the training phase, using the Baum-Welch algorithm [9]. The unknown patterns that have to be tested in the training phase are evaluated by every model. For every HMM we get a likelihood for the tested observation sequence **O**. The HMM which has the highest likelihood determines the class of the unknown pattern, see Fig. 2. In our approach we use a different procedure, training only one HMM model for all voice qualities, see section 4.4.

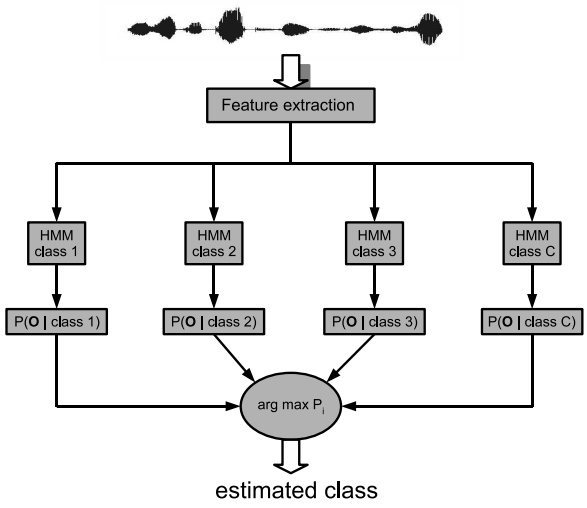

Figure 2: Classifying with HMM

4.3. Classifying voice qualities

We apply the classification with HMM to the problem of voice quality recognition. For this application a left-to-right model as used in ASR is not suitable, since a special voice quality does not occur at a specific time during an

utterance. Thus, we use the ergodic model shown in Fig. 3. Here, all state transition are allowed. The number of states is equal to the number of voice qualities contained in the database which is 4 in our case. For every speech segment of an utterance the 6-dimensional feature vector $\underline{x}(t)$ is extracted. With the help of the vector quantization, for every feature vector $\underline{x}(t)$ the corresponding symbol $o(t)$ is obtained. Therefore the k-means algorithm is applied, using 100 iterations. For one of the 4 states j one of 16 observations k with probability $b_j(k)$ can be emitted.

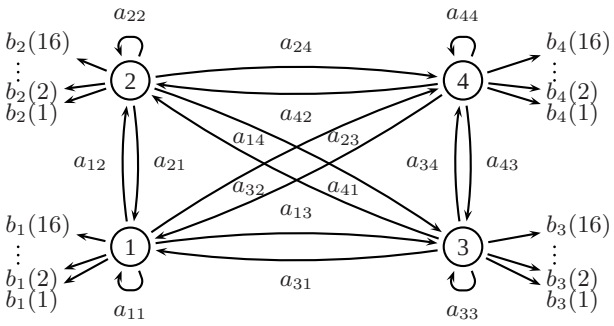

Figure 3: Ergodic HMM with four states and 16 symbols

4.4. Extraction of voice quality contours

Training an HMM for every voice quality would result in a classification of a whole utterance to one voice quality. This implies that the adopted states can only be compared within one model. But we aim for a classification on the segmental level. Since the voice quality is not fixed during an utterance the classification of single analysis segments is more relevant than the classification of whole utterances. Therefore only one HMM is trained with the help of the Viterbi algorithm which considers only the most likely state sequence. That means for every time instant the HMM adopts one state that can be interpreted. Although the states are hidden, they can have a physical meaning. We found out that, in analogy to ASR where a state corresponds to a phoneme, for voice quality recognition a state can be associated with a voice quality.

4.5. Associating voice qualities with HMM states

The most likely state sequence for the training utterances is generated using the Viterbi algorithm and a histogram for each voice quality is extracted. By searching the most frequently state for one voice quality a unique mapping between the state and the voice quality is established. Table 1 shows an example. We can see that for creaky voice in 81.5% of the cases state 4 occurs. Thus, the row for creaky and also the column for state 4 are deleted from the matrix since there is now a mapping. For the remaining matrix again the maximum value is searched, the relationship is built, and the corresponding row and column are deleted. This procedure is repeated until the matrix is empty and all states are mapped to a voice quality.

vq / state	state 1	state 2	state 3	state 4
modal	0.083	0.210	**0.697**	0.010
breathy	0.181	**0.707**	0.108	0.004
creaky	0.136	0.034	0.015	**0.815**
rough	**0.578**	0.100	0.266	0.056

Table 1: Analysis of voice qualities and states

5. Results

In this section results for the segmental classification of voice qualities are presented. First, the performance of our approach using a discrete 4-state HMM is compared with a clustering using the well known k-means algorithm. Then, two exemplary contours of utterances with changing voice quality are plotted.

5.1. K-MEANS vs. HMM

There are only very few databases available that contain non-modal voice qualities to a greater extent. For our first study 563 utterances, with over 58000 segments, spoken by four male speakers are used. Each utterance contains only one of the four voice qualities.

Here, the speaker dependent classification of voice qualities on a segmental level is studied. Therefore, our proposed approach using HMM is compared to the k-means algorithm. To get the reference class for each segment, we assumed that every segment of the utterance is spoken in the same voice quality. Thus, all segments of a sentence are labelled with the same voice quality. Table 2 shows the confusion matrix of segmental voice quality recognition using the k-means algorithm and HMM.

K-MEANS	modal	breathy	creaky	rough
modal	**37.9**	9.5	38.4	14.2
breathy	28.6	**42.8**	11.6	17.0
creaky	13.7	11.4	**64.3**	10.6
rough	24.1	16.7	34.5	**24.7**
HMM	**modal**	**breathy**	**creaky**	**rough**
modal	**38.9**	9.2	17.4	34.5
breathy	18.4	**56.9**	5.5	19.2
creaky	7.3	12.8	**77.1**	2.8
rough	15.0	15.1	8.6	**61.3**

Table 2: Segmental voice quality classification

As we see, our proposed 4-state HMM outperforms the k-means algorithm for all voice qualities. The greatest difference between k-means and HMM is observed for rough voice with 36%. For k-means, rough voice is mostly classified to creaky voice, whereas HMM is right for 61.3% of the segments. The overall classification rate is improved by 11.4%.

5.2. Voice quality contours for mixed voice qualities

For extracting the contours further utterances spoken by the same speakers with two changing voice qualities are used. The first part is uttered non-modal while the second part is spoken in modal voice.

Two voice quality contours are exemplary extracted using our proposed 4-state HMM. Every voiced speech segment, for which the feature vector could be calculated is assigned to one voice quality. The remaining segments are rejected as invalid. In Fig. 4, the contour of a creaky and modal voice is depicted above the corresponding waveform. For the first part the contour adopts solely the state that can be associated with creaky voice. For the second part mainly modal voice is recognized.

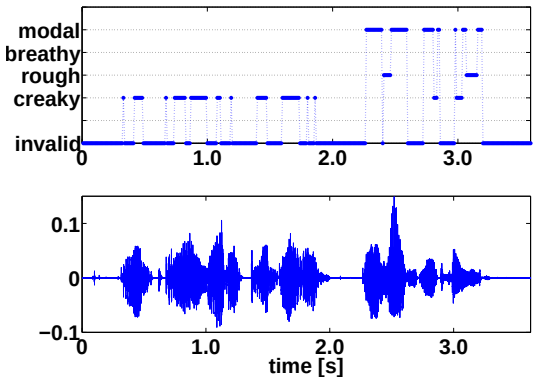

Figure 4: Voice quality contour of a first creaky and then modal voice

The first part of the second utterance is spoken in breathy voice, the second part contains mainly modal voice quality. Fig. 5 shows the extracted voice quality contour. One can see, that for all valid segments in the first part breathy voice is classified. For the second part a mix of mainly modal and creaky voice is recognized. A few breathy and rough segments are also detected.

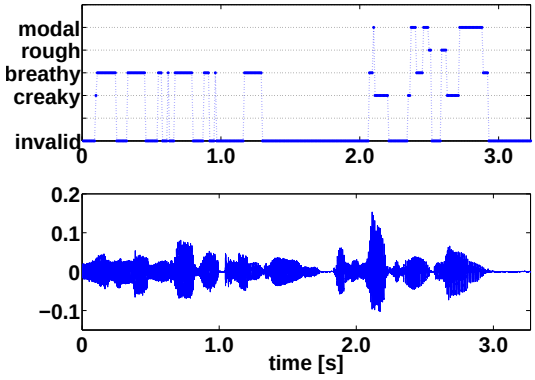

Figure 5: Voice quality contour of a first breathy and then modal voice

6. Conclusions

In this paper we presented a method for extracting voice quality contours using discrete hidden Markov models. As feature vector a 6-dimensional voice quality parameter vector is used. Thereby, the Viterbi algorithm is used to generate the most likely state sequence of the trained model. Every state of the HMM can be interpreted as a discrete voice quality defined by Laver. So each segment can be associated with a voice quality. Doing so for every segment of a spoken utterance a voice quality contour is extracted. We showed that our 4-state HMM outperforms the k-means clustering by over 11% on the segmental level. The generated voice quality contours coincide with the perceived voice quality for the utterances recorded in an anechoic room.

7. References

[1] John Laver, *The phonetic description of voice quality*, Cambridge University Press, 1980.

[2] W. Wokurek and M. Pützer, "Automated corpus based spectral measurement of voice quality parameters," *Proceedings of the International Congress of Phonetic Sciences, Barcelona*, pp. 2173–2176, 2003.

[3] M. Masarek and M. Pützer, "Differenzierung gesunder Stimmqualitäten und Stimmqualitäten bei Rekurrensparese mit Hilfe elektroglottographischer Messung und RBH-System," *Sprache Stimme Gehör*, 2000.

[4] K. Classen, G. Dogil, M. Jessen, K. Masarek, and W. Wokurek, "Stimmqualität und Wortbetonung im Deutschen," *Linguistische Berichte*, vol. 174, pp. 202–245, 1998.

[5] M. Lugger, B. Yang, and W. Wokurek, "Robust estimation of voice quality parameters under real world disturbances," *Proc. IEEE ICASSP*, 2006.

[6] Francis Nolan, *A figure of speech*, chapter Forensic speaker identification and the phonetic description of voice quality, pp. 385–411, London: Lawrence Erlbaum Associates, 2005.

[7] M. Lugger and B. Yang, "Classification of different speaking groups by means of voice quality parameters," *ITG-Sprach-Kommunikation*, 2006.

[8] K. Stevens and H. Hanson, "Classification of glottal vibration from acoustic measurements," *Vocal Fold Physiology*, pp. 147–170, 1994.

[9] L.R. Rabiner, *Fundamentals of speech recognition*, Prentice Hall, 1993.