EVALUATION OF SOUND CLASSIFICATION ALGORITHMS FOR HEARING AID APPLICATIONS

JuanJuan Xiang, Martin F. McKinney, Kelly Fitz, Tao Zhang

Starkey Laboratories, 6600 Washington Avenue South, Eden Prairie, MN, 55344
Email: juanjuan_xiang@starkey.com

ABSTRACT

Automatic program switching has been shown to be greatly beneficial for hearing aid users. This feature is mediated by a sound classification system, which is traditionally implemented using simple features and heuristic classification schemes, resulting in an unsatisfactory performance in complex auditory scenarios. In this study, a number of experiments are conducted to systematically assess the impact of more sophisticated classifiers and features on automatic acoustic environment classification performance. The results show that advanced classifiers, such as Hidden Markov Model (HMM) or Gaussian Mixture Model (GMM), greatly improve classification performance over simple classifiers. This change does not require a great increase of computational complexity, provided that a suitable number (5 to 7) of low-level features are carefully chosen. These findings indicate that advanced classifiers can be feasible in hearing aid applications.

Index Terms **—** sound classification, hearing aids, Hidden Markov Model, Gaussian classifiers, feature selection

1. INTRODUCTION

Hearing aid users are typically exposed to a variety of listening situations, such as speech, music and noisy environments. To yield the best listening experience, the behavior of the instrument, for instance the activation of a directional microphone or the compression/expansion parameters, should adapt to the currently engaged environment. This indicates the need for sound classification algorithms functioning as a front end to the rest of the signal processing scheme housed in the instruments [1].

Sound classification has been studied under different contexts, such as speech/music discrimination [2, 3], environment sound classification [4], and content-based audio classification [5, 6]. Compared with these applications, sound classification in hearing aids is more challenging due to the limited power consumption, the real time operation and the great varieties of sound encountered in the real life. So far, a couple of simple features and classifier schemes, such as a threshold-based classifier, have been implemented in hearing aids to identify speech, noise and speech in noise [7] . When more kinds of sounds need to be classified, advanced classifiers and features might have to be involved to achieve satisfactory performance. The goal of this study is to systemically evaluate the impact of sophisticated features and classifiers on the classification rate, computational cost and classification delay. A classification system which is intended to detect speech, music and several kinds of noises is constructed. The performance of two feature sets, low-level features and Mel-scale frequency cepstral coefficients (MFCC), are compared by applying Gaussian classifiers, GMM and HMM individually.

2. METHODS

A two-stage environment classification scheme is implemented in this study. The signal is first classified as music, speech or non-speech. Then the non-speech sounds are further characterized as machine noise, wind noise or other sounds. At each stage, the classification performance and the associated computational cost are evaluated along three dimensions: the choice of classifier, the choice of feature set, and number of features within each feature set. Each component is described in detail in the following sections.

2.1. Audio Features

Choosing appropriate features is a domain-specific question. Based on previous work [1-3, 8], two feature groups are investigated in this study, specifically a low-level feature set, and MFCCs. The former consists of both temporal and spectral features, such as zero crossing rate, short time energy, spectral centroid, spectral bandwidth, spectral roll-off, spectral flux, high/low energy ratio, etc. The logarithms of these features are included in the set as well. The first 12 coefficients are included in the MFCC set [9]. There are some other features proposed in literature, such as cepstral modulation ratio [10] and several psychoacoustic features

[8, 11]. These features are not investigated here either due to their high computational cost or because the calculation of these features is not well defined.

Within each set, some features may be redundant or noisy or simply have weak discriminative power. To identify optimal features, a forward sequential feature selection algorithm is employed [12]. It's noteworthy that the derived feature set is specific to the choice of classifiers, which are discussed in the following section.

2.2 Classifier selection

In the past years many pattern-recognition techniques have been proposed and used in various fields. However, for hearing aid applications, it is crucial to keep computational cost low. For this purpose, this study focuses on three classification algorithms: a quadratic Gaussian classifier, a GMM with 5 components, and an ergodic HMM with 5 states and 5 components [13]. The feature selection algorithm is performed for each classifier. The training of GMM and HMM is carried out using the expectation-maximization (EM) algorithm, and classification decisions for the HMM are based on the Viterbi decoder method [14]. To examine the robustness of performance for a given combination of classifiers and features, a 4-fold cross-validation testing procedure is employed to determine the average classification error rate [13].

2.3 Audio database

The evaluation of the investigated features and classifiers is performed on a database composed of sounds from five classes: music, speech, wind noise, machine noise and others. The music content is taken from a database collected by Scheirer and Slaney [3], which contains 80 15-second audio music samples covering different genres, such as classical music, rock and pop songs, folk music, etc. The remaining samples are recordings made by the authors of the current study. The acoustic signals from a variety of auditory scenes were picked up by a microphone located in a BTE hearing aid first and then are stored in a Sony TCD-D8 DAT recorder with a 16-bit resolution and a 48 kHz sampling frequency. The recordings were manually divided according to recording environment and then segmented using a short-term energy detector, followed by manual verification and adjustment of the segment boundaries. The resulting segments were used for training and testing the classification system.

The class "speech" includes both clean and noisy speech. The clean speech comprises of speech spoken by different people in different reverberation situations, such as a living room or a cafeteria. The noisy speech is generated by randomly mixing selected files from the clean speech class with noise at three levels of SNR: -6dB, 0dB and 6 dB. The class "machine noise" contains the noise generated by various machines, such as automobile, vacuum and blender. The class "others" is the most varied category comprising any sounds that are not suitably described by the other three classes, for instance the sounds from water running, foot stepping, etc. The duration of the samples of each class is listed in table 1.

Sound Type	Music	Speech	Machine noise	Wind noise	Others
Duration (minutes)	14	40	73	12	22

Table 1: List of the recorded sound types and their length of durations.

2.4 Computational cost

Computational cost is a critical constraint concerning the application of various classification algorithms in hearing aids, mainly due to the limitation on power consumption and real time operation. The cost of a classification algorithm consists of two parts, feature evaluation and classification. The former is related to the length of analysis window and the resolution of the Fourier Frequency Transform (FFT), while the latter is mainly determined by the number of sound classes and the dimensionality of the employed feature vector. For a GMM and HMM classifier, the number of components and states affect the computational cost as well. At each classification stage, the computational cost is measured in terms of number of operations and evaluated along three dimensions: choice of classifier, choice of feature set, and number of selected features, just as in the performance evaluation.

3. EVALUATION RESULTS

3.1 Choice of classifiers and feature sets

The evaluation performed at each stage is combined to obtain the overall classification error rate, which is illustrated in Fig. 1. The plot shows the error rate obtained at each iteration of the forward sequential feature selection process. The various classifiers and feature sets are indicated by line styles and marker styles, respectively.

Several results are apparent upon examination of Fig. 1. The first is that advanced classifiers perform better on average than a simple one. Specifically, when ten features are employed, the lowest error rate of the Gaussian classifier is 26%, while for the GMM and HMM, the rates are 18% and 12%, respectively. The performance improvement associated with the employment of GMM might be explained by the better fitting between the distribution of feature vectors and the model. And the further improvement of HMM might be related with its exploitation of the dynamics of the feature vectors.

In terms of the feature set, we observe that there is no significant difference in classification performance between

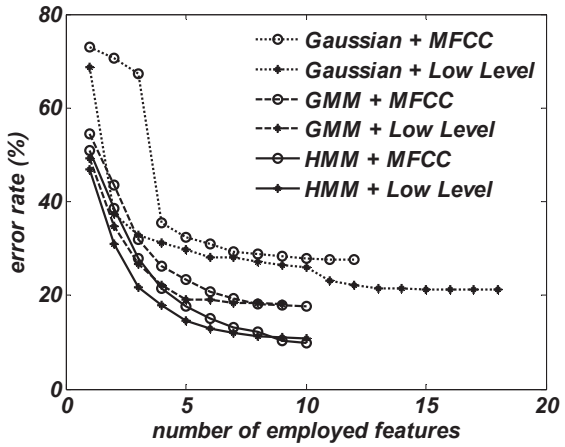

Figure 1: Error rate as a function of the number of employed features. Performance is evaluated over two feature sets (MFCC and low level) and three models (Gaussian, GMM and HMM).

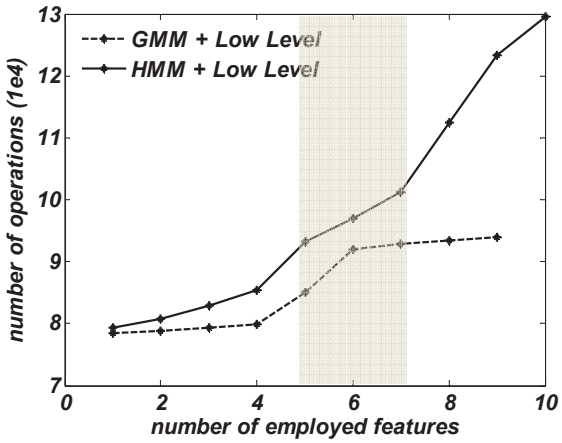

Figure 2: Computational cost as a function of the number of employed features. Performance is evaluated over advanced models (GMM and HMM) and low level feature set. The recommended number of features is indicated by the gray patch.

the low level feature set and MFCC, provided that the number of employed features is more than five in both cases. This demonstrates that a carefully chosen low-level feature subset has the same discriminative power as the MFCC set. Considering that the computational cost of low-level feature extraction is typically one order of magnitude lower than a MFCC extraction, the low-level features are a better choice when computational resources are limited.

3.2 Optimal number of features

From the above discussion the advantage of using advanced classification models with the low-level feature set becomes obvious. Therefore, the following discussion focuses on the advanced classifiers with the low-level feature set. In this section we examine the impact of the number of features employed. The overall computational cost is determined from the two stages (Fig. 2). It is noteworthy that when combining the computational cost from the two stages, some features are identified as optimal features on both stages but need only be calculated once. Thus the overall cost is less than the direct summation over the two stages.

A comparison between Fig. 1 and Fig. 2 shows that the increased number of features results in both decreased error rates and increased computational costs, thus indicating a trade-off between performance and computational complexity. It seems that choosing five to seven features is a reasonable comprise between the two factors. Using this number of features, the error rate is about 10% and the computational cost is still manageable. On the other hand, using more than seven features only slightly improves the performance but incurs great computational cost.

3.3 Recognition rate as a function of test sequence length

In this section we further examine the impact of the test sequence length on performance when using low-level features and advanced classifiers. The error rates are plotted as function of the test sequence length in Fig. 3. As expected, increasing the length of test sequence improves the classification performance. An approximate 20% decrease of error rate is obtained by increasing the test sequence from 128ms to 256ms. This benefit is diminished with further increase in the length of the test sequence. The overall pattern of the rate of decrease seems to be consistent across classifiers.

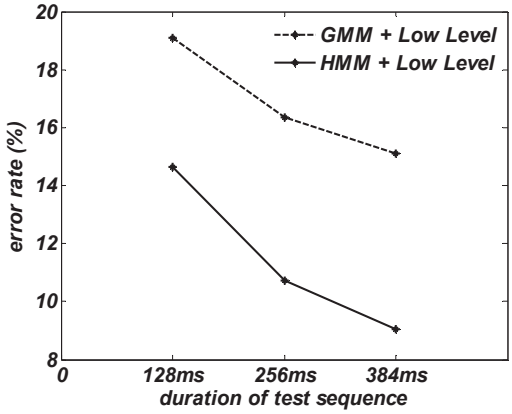

Figure 3: Error rate as a function of the length of test sequence. The performance is evaluated over advanced models and low level feature set.

3.4 Recognition rate as a function of classification structure

Finally we compare the recognition rates of advance classifiers using a two-stage classification scheme with the one based on a flat structure where the input sound is directly assigned to one of the five types. For each case the low-level features are employed in the feature selection process and the lowest error rates are presented (Fig. 4). It

seems that the flat classification scheme has a slight advantage over the two-stage version, at the expense of flexibility and computational complexity.

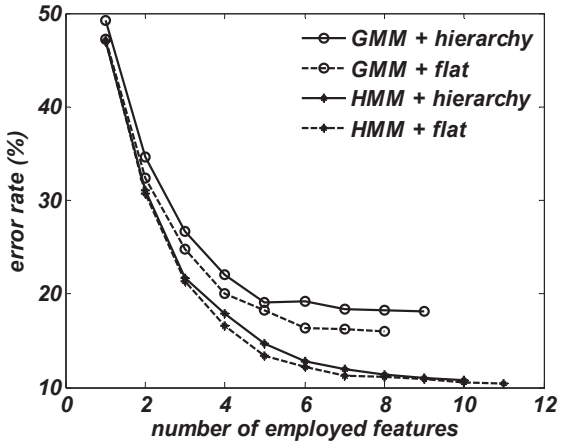

Figure 4: Error rate as a function of the number of employed features. Performance is evaluated over two classification structures (a hierarchy one and a flat one), two classifiers (GMM and HMM) and the low-level feature set is used.

4. CONCLUSION

A number of experiments are conducted to assess the impact of classifiers, feature sets and number of features on the performance of classification systems, where five sound classes, "speech", "music", "machine noise", "wind noise" and "others", are distinguished. The results show that compared with a Gaussian classifier, advanced models, such as GMM or HMM, greatly improve the classification performance. The use of the advanced classifiers is not necessarily associated with a great increase of computational complexity, as one may expect. As for the choice of feature set, the performance of low-level-feature-based classification is comparable with MFCC-based classification. Considering that the computational cost of low-level features is generally lower than MFCC, the low-level feature set should be recommended when the computational resource is limited. In addition, the number of features is suggested as 5 ~ 7 to balance the performance and computational cost. The classification performance can be further improved by using longer test sequence or a flat classification scheme.

Future work will involve the refinement of the proposed system, such as employing the optimal number of states and components for different sound types, enlarging the audio database so that the evaluation is more robust and further identifying different kinds of environment sounds belonging to the "others" sound class.

REFERENCES

[1] J.M. Kates, "Classification of background noises for hearing aid applications," *Journal of the Acoustical Society of America*, **97**(1): pp. 461-470, 1995.

[2] Y. Lavner and D. Ruinskiy, "A Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation," *EURASIP Journal on Audio, Speech, and Music Processing*, 2009. doi:10.1155/2009/239892

[3] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1331-1334, 1997.

[4] S. Chu, S. Narayanan, and C.C.J. Kuo, "Environmental sound recognition using MP-based features," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1-4, 2008.

[5] R. Huang and J.H.L. Hansen, "Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, **14**(3): pp. 907-919, 2006.

[6] T. Zhang and C.C.J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Transactions on Speech and Audio Processing*, **9**(4): pp. 441 - 457, 2001.

[7] B.W. Edwards, Z. Hou, C.J. Struck, and P. Dharan, "Signal-processing algorithms for a new software-based, digital hearing device," *Hearing Journal*, **51**(9): pp. 44-54, 1998.

[8] M.F. McKinney and J. Breebaart, "Features for audio and music classification," in *Proceedings of International Conference on Music Information Retrieval*, pp. 151-158, 2003.

[9] T.F. Quatieri, *Discrete-time speech signal processing*, Prentice Hall PTR, 2002.

[10] R. Martin and A. Nagathil, "Cepstral modulation ratio regression (CMRARE) parameters for audio signal analysis and classification," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 321-324, 2009.

[11] M. Buchler, S. Allegro, S. Launer, and N. Dillier, "Sound classification in hearing aids inspired by auditory scene analysis," *EURASIP Journal on Applied Signal Processing*, **2005**(18): pp. 2991-3002, 2005.

[12] A.L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial intelligence*, **97**(1-2): pp. 245-271, 1997.

[13] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern classification*, Wiley New York, 2001.

[14] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, **77**(2): pp. 257-286, 1989.