Proceedings of the 10th International Conference on Music Perception and Cognition (ICMPC 10). Sapporo, Japan.
Ken'ichi Miyazaki, Yuzuru Hiraga, Mayumi Adachi, Yoshitaka Nakajima, and Minoru Tsuzaki (Editors)

Analysis and Automatic Detection of Breath Sounds in Unaccompanied Singing Voice

Tomoyasu Nakano[†1,*] Jun Ogata[‡2] Masataka Goto[‡3] Yuzuru Hiraga[†4]

[†]*Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan*
[‡]*National Institute of Advanced Industrial Science and Technology (AIST), Japan*
[1]t.nakano@aist.go.jp, [2]jun.ogata@aist.go.jp, [3]m.goto@aist.go.jp, [4]hiraga@slis.tsukuba.ac.jp

ABSTRACT

This paper presents a dual approach to the study of breath sounds in singing, consisting of an acoustic analysis of breath sounds, and development of an automatic breath detection system. Previous work on automatic breath detection were based on relatively simple features that were postulated to be relevant to the detection. In contrast, this study starts with a detailed acoustic analysis of breath sounds, with the aim to explore novel characteristics. The obtained results can be used to enhance the capability of automatic breath detection.

The acoustic analysis used singing voice recordings of 18 singers with a total length of 128 mins (1488 breath events). The results of the analysis show that the spectral envelope of breath sounds remain similar within the same song, and their long-term average spectra have a notable spectral peak at about 1.6kHz for male singers and 1.7kHz for female singers. A prototype version of a breath detection system was implemented, using HMM based on MFCC, ΔMFCC, and Δpower as acoustic features. In an evaluation experiment with 27 unaccompanied song samples, the system achieved an overall recall/precision rate of 97.5%/77.7% for breath sound detection.

I. INTRODUCTION

All singing contains breath sounds of the singers, regardless of whether we consciously recognize them or not. Analysis and automatic detection of breath sounds is becoming a vital research topic, leading to various applications. Detected breath sounds can be suppressed as noise, but can also be used as valuable cues for applications including segmentation and structural analysis of music, and evaluation of the singer's skills.

Ruinskiy & Lavner [1] present an automatic breath detection method of speech/song signals. Their method was based on a template matching approach using MFCCs (Mel-Frequency Cepstrum Coefficients), short-time energy, ZCR (zero-crossing rate), spectral slope, and duration as key features. They also introduce an edge detection algorithm to extract the boundaries of each breath event. In an evaluation experiment using speech/singing voice recordings of 22 singers and 2 narrators with a total length of 24 minutes (334 breath events), their method achieved a recall/precision rate of 97.6%/95.7% for the highest-tuned setting.

Price *et al.* [2] and Wightman & Ostendorf [3] both present automatic breath detection methods of speech signals based on cepstral coefficients. Price *et al.* [2] used GMM (Gaussian Mixture Model) as the classifier, and achieved a detection rate of 93%. Wightman & Ostendorf [3] used a Baysian classifier and achieved a detection rate of 91.3% for the highest-tuned setting.

These work did not attempt an in-depth analysis of breath sounds, and the features used for detection were relatively simple. A detailed analysis of the breath sound data may reveal characteristic features that are interesting in their own right, and also are valuable for improving and enhancing the capabilities of breath detection systems.

So this study takes to a dual, complementary approach consisting of a detailed acoustic analysis of breath sounds, and development of an automatic breath detection system. The results of the analysis can be utilized and tested on the detection system, while the performance of the system will in turn enhance our knowledge about breath sounds and suggest further issues to be investigated.

Section II. describes the breath sound analysis. The analysis is focused on the spectral features of the singing voice. Section III. describes a prototype implementation of our detection system based on HMM (Hidden Markov Models), and the results of a preliminary evaluation experiment.

The singing sample data used in this study are 16bit/16kHz/monaural digital recordings of unaccompanied singing, either sung *a capella*, or vocal tracks extracted from a full accompanied performance.

II. ANALYSIS OF BREATH SOUNDS

Previous work [1, 2, 3] have shown that cepstrum-based acoustic features are effective for breath sound detection. This suggests that characteristic features of breath sounds exist in their spectral envelope, which is the focus of the analysis described below.

A. The Song Dataset

Song samples are taken from the *RWC Music Database: Popular Music* (RWC-MDB-P-2001) [4] and the *AIST Humming Database* (AIST-HDB) [5]. RWC-MDB contains song music sung by professional singers, with lyrics in either Japanese or English. AIST-HDB contains singing recited by 100 subjects upon hearing songs from the RWC Music Database (Popular Music [4] and Music Genre [6]). Most of the subjects were non-musicians and had no extensive training on singing.

The analysis used 27 songs by 16 singers from RWC-MDB and 100 samples by 2 subjects from AIST-HDB. The AIST-HDB samples are 50 excerpts from 25 songs in English from RWC-MDB. In the analysis, these excerpts are merged into a single entry for each subject. So the entire dataset has 29 entries (27 songs and 2 merged excerpts), of which 15 were by male singers and 14 by female singers (see Table 1 for the entry list). The total length of the songs is 128 minutes.

B. Method

Breath events used in the analysis were extracted from the song samples by hand-marking. The spectral envelopes of the extracted events were calculated by using the cepstrum method, and their long-term average were also obtained. The spectra were calculated by using short-term Fourier transform shifted by 10.0 ms (160 points) with a 1024-point (64.0 ms) Hanning window. The first three formant frequencies (F1, F2, and F3) were extracted by using WaveSurfer [7].

*Presently, with National Institute of Advanced Industrial Science and Technology (AIST), Japan.

ISBN: 978-4-9904208-0-2

Table 1. **The dataset with total length of 128 minutes used in the analysis and detection experiment.**

RWC-MDB-P-2001

Piece no.	Singer name	Gender	Lyrics language	Song length (sec)	Number of breath events	Cumulative length of breath events (sec)	Statistics of breath events (msec)			
							mean	S.D.	min	max
No.001	Kazuo Nishi	Male	Japanese	207.2	54	9.7	179.7	47.7	81.7	311.6
No.009	Kazuo Nishi	Male	Japanese	275.0	54	13.7	253.1	85.6	100.0	432.5
No.012	Kazuo Nishi	Male	Japanese	202.8	45	10.3	228.6	107.7	55.0	530.0
No.004	Hisayoshi Kazato	Male	Japanese	240.5	10	2.0	199.5	79.1	105.0	325.0
No.011	Hisayoshi Kazato	Male	Japanese	265.9	23	7.0	304.0	134.7	137.5	535.0
No.019	Hisayoshi Kazato	Male	Japanese	287.0	31	5.4	174.0	63.4	87.5	412.5
No.006	Oriken	Male	Japanese	204.5	43	16.4	380.6	187.5	105.0	975.0
No.015	Katsuyuki Ozawa	Male	Japanese	160.9	9	2.5	277.8	222.2	102.5	832.5
No.037	Yoshinori Hatae	Male	Japanese	237.1	59	15.2	257.5	68.7	157.3	532.8
No.038	Kousuke Morimoto	Male	Japanese	273.3	49	13.5	275.6	96.8	95.0	527.5
No.048	Hiroshi Sekiya	Male	Japanese	269.6	48	22.0	457.4	208.4	190.0	1225.0
No.057	Masashi Hashimoto	Male	Japanese	265.1	62	23.0	371.2	157.8	175.0	845.0
No.085	Jeff Manning	Male	English	208.5	53	11.2	212.2	60.2	95.0	375.0
No.095	Jeff Manning	Male	English	230.8	70	20.8	296.8	185.8	100.9	1074.0
No.100	Shinya Iguchi	Male	English	293.3	39	19.0	487.4	168.0	199.6	932.5
No.007	Tomomi Ogata	Female	Japanese	296.7	12	3.1	262.2	94.2	124.3	442.8
No.018	Tomomi Ogata	Female	Japanese	252.1	48	9.3	193.8	73.2	50.0	347.5
No.014	Rin	Female	Japanese	232.9	60	14.5	242.2	88.6	75.0	585.0
No.021	Rin	Female	Japanese	266.9	60	20.5	342.0	128.2	107.5	590.0
No.016	Hiromi Yoshii	Female	Japanese	262.1	43	10.9	254.0	108.5	85.0	497.5
No.017	Hiromi Yoshii	Female	Japanese	239.4	55	16.1	292.3	110.8	122.5	687.5
No.075	Hiromi Yoshii	Female	Japanese	199.9	33	10.7	323.7	98.5	123.2	542.7
No.077	Makiko Hattori	Female	Japanese	234.1	51	15.4	301.3	126.6	135.0	757.5
No.092	Betty	Female	English	218.4	37	11.3	304.4	108.3	122.5	690.0
No.094	Betty	Female	English	221.9	44	11.5	261.1	92.5	112.5	452.5
No.091	Donna Burke	Female	English	221.8	48	15.1	315.0	109.2	117.5	690.0
No.097	Donna Burke	Female	English	241.1	63	18.1	287.1	158.0	66.1	786.1

AIST-HDB

Singing skill	Singer's ID	Gender	Lyrics language	Entry length (sec)	Number of breath events	Cumulative length of breath events (sec)	Statistics of breath events (msec)			
							mean	S.D.	min	max
poor	E001	Female	English	595.9	144	45.9	318.5	143.5	87.5	807.5
good	E008	Female	English	558.6	141	44.5	315.3	160.6	102.5	1069.3

C. Results

The entries of the dataset, together with their breath event data, are shown in Table 1. The breath event data items are the following.

- song/entry length
- number of breath events
- cumulative length of breath events
- breath event statistics (mean, standard deviation, min and max)

The "Singing skill" column for AIST-HDB entries show a binary rating (good/poor) obtained from a rating experiment judged by human subjects [8]. By this rating, all singers of the RWC-MDB can be considered as "good".

The total number of extracted breath events is 1488 with an overall average of 51.3 (44.6 for RWC-MDB entries, and 142.5 for AIST-HDB entries). The total cumulative length of breath events is 438.5 sec., corresponding to 5.7 % of the total sample length (128 mins.). Breath event length is widely ranged from 50ms. (min.) to 1225ms. (max.).

Figure 1 shows example analysis results from two songs, No.038 (male singer, Japanese lyrics) and No.097 (female singer, English lyrics). The spectrograms to the left show the data for breath events, extracted from the entire song samples (removing the vocal and silent parts). The graphs to the right show the long-term average of the spectral envelopes.

Both spectrograms show that the spectral pattern of breath events are stable throughout the song, and there are significant spectral peaks, most prominent at the F2 frequency. These peaks are retained in the long-term average spectra as well.

Figure 2 shows the long-term average spectra of breath sounds for all 29 entries. The entries by male singers are gathered in the upper half of the figure, and those by female singers are gathered in the lower half. Individual singers are identified by their names (ID numbers for AIST-HDB entries) to the left of the graphs.

From the figure, it can be seen that the spectral pattern is mostly identical for the same individual singer. And although the patterns may differ among individuals, a notable point is that there is a significant peak at a common frequency across different individuals (see below).

D. Discussion

Summarizing the above, the results suggest that the spectral envelopes of breath sounds remain similar within the same song (Figure 1), and their long-term average spectra have the following characteristics (Figure 2):

(1) Spectral envelope for the same singer remains invariant across different songs,
(2) There is a common, significant spectral peak at about 1.6kHz (for male singers) and 1.7kHz (for female singers),
(3) Secondary peaks exist in the range of 850Hz–1kHz, more prominent in the female singers.

The existence of common spectral peaks (with a slight difference according to gender) suggests the existence of a common cause. This may be due to breathing mechanisms that are involuntary and invariant to the body size, voice range, or singing context.

There are, however, cases where the peaks are not that significant (entries No.048 (male), No.007, No.014, No.021 (female)). Looking into the temporal spectral data of such cases, the spectral peaks fluctuated widely within the range of 1.5kHz to 3kHz. But in these cases, the secondary peaks in the range of 850Hz–1kHz

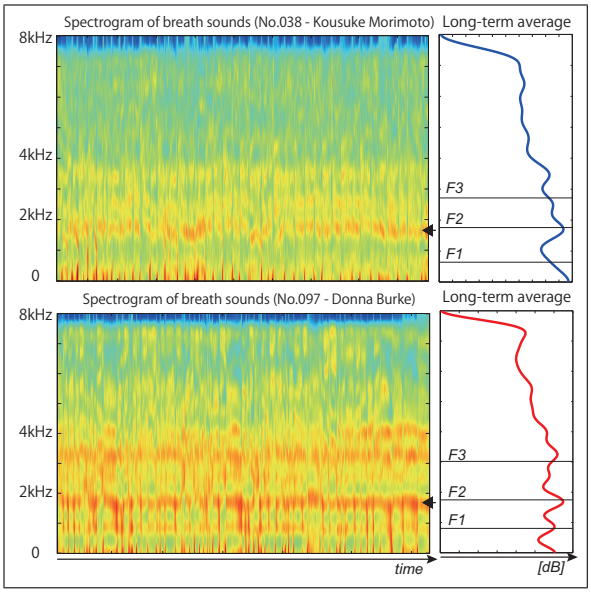

Figure 1. Example spectrograms of breath sounds (left diagram) and their long-term average (right graph) taken from entries No.038 (top) and No.097 (bottom). The average of the formant frequencies (F1, F2, and F3) are shown in the right graph.

were relatively stable and significant, so this may be a reflection of certain singing methods or techniques.

As seen from the above, breath events have a relatively stable and context-free nature that can be captured by data in the spectral envelope domain. These results provide grounding for why previous breath detection methods [1, 2, 3] based on cepstrum data were effective and achieved high detection rates.

A point of note is that the extracted length of breath events are wide-ranged (50msec–1225msec: see Table 1), possibly due to the nature of the music. This means that a breath detection system must deal with such variation, *e.g.* recognizing continuation of a long, single breath event.

III. DETECTION OF BREATH SOUNDS

Based on the results of the analysis described above, a prototype automatic breath detection system was implemented, and was tested in a preliminary evaluation experiment.

A. System Outline

The system uses HMM (Hidden Markov Model) as its base framework, and is a composite of three HMM's. Each of the HMM's were dedicated to detect whether the current data is a breath sound, a singing voice, or a silent section. The combined results of the HMM's become the overall recognition result of the system. The HMM's used MFCCs (Mel Frequency Cepstral Coefficients), ΔMFCCs and Δpower as feature vectors. This selection follows the significance of the spectral domain obtained from the analysis results.

The reason for using HMM's is because they have a preferrable nature of being able to deal with variant time-length events, and are able to track data variance along time. Another merit is that they can be used in combination with detection of other features such as lyrics recognition.

In the implementation, the Hidden Markov Model Toolkit (HTK) [9] was used for feature extraction, Viterbi alignment and training of the models. Table 2 shows the system setting for the evaluation experiment described below.

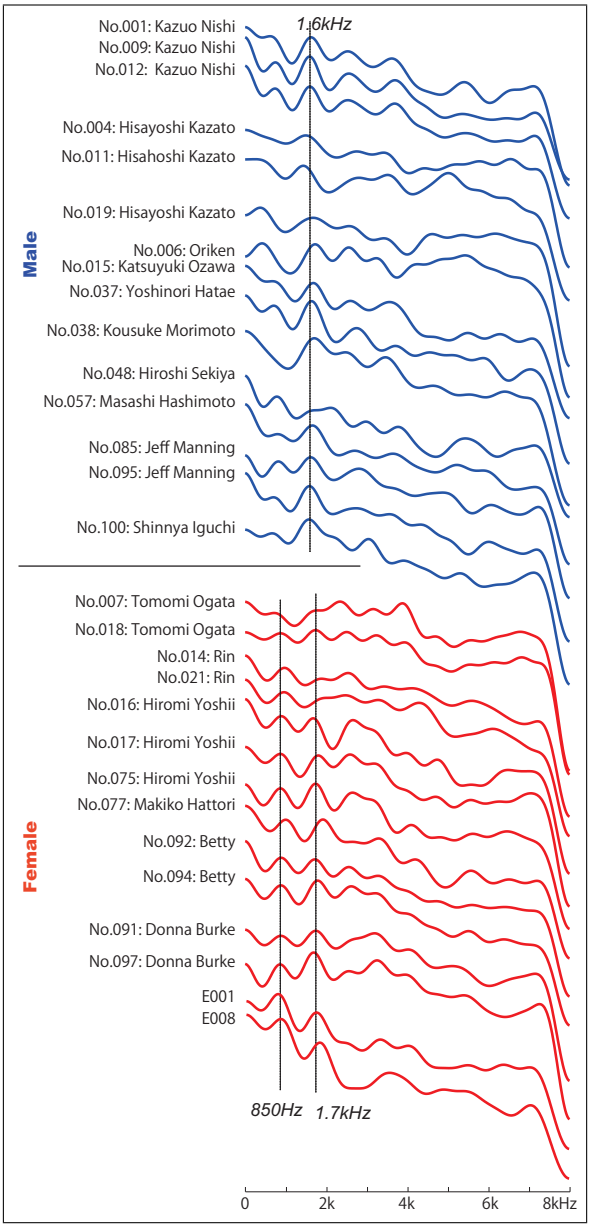

Figure 2. The long-term average spectra of the breath sounds and their significant peak frequencies. The entries for male singers are gathered in the upper half, and female singers, lower half.

B. Evaluation Experiment

The detection system was evaluated on the 27 song samples from the RWC Music Database (RWC-MDB-P-2001 in Table 1).

The system was evaluated using the cross validation method. Cross validation reserves one sample as the test data, and the other data are used as training data for the system. The trained system is then evaluated on the reserved test data. This procedure is performed for each of the 27 songs, so the results show the system performance on a single song (test data) by a system trained on other data. In the actual experiment, not only the test data itself, but also songs by the same singer were excluded from the training data.

System performance is evaluated by recall (R) and precision (P) rates, defined as follows.

$$R = \frac{\text{events correctly detected as breath}}{\text{events in breath}} \times 100 \quad (1)$$

$$P = \frac{\text{events correctly detected as breath}}{\text{events detected as breath}} \times 100 \quad (2)$$

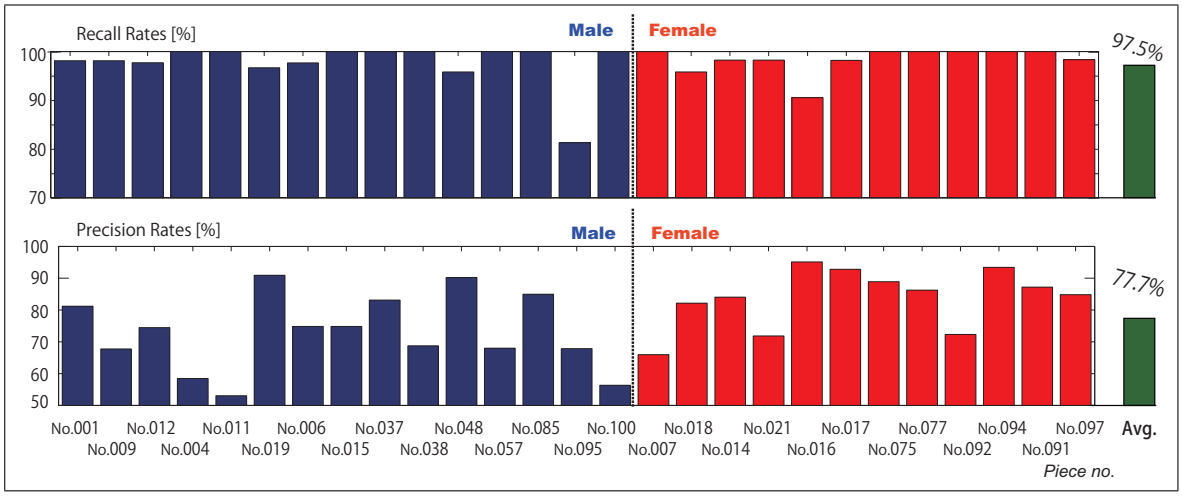

Figure 3. Result of the evaluation experiment (Upper histogram: Recall rates; Lower histogram: Precision rates).

Table 2. System setting for the breath detection experiment.

Sampling rate	16kHz
Window function	Hamming
Frame length	25ms
Frame period	10ms
Feature vector	12th order MFCC 12th order ΔMFCC Δ Power
HMM	3 states, 16 mixtures

In short, recall rates show the percentage of breath events detected by the system, while precision rates show the percentage of correct detections made by the system.

C. Results & Discussion

Figure 3 shows the results of the evaluation experiment in terms of recall/precision rates for each of the songs.

The overall (average) recall rate was 97.5%, while the overall precision rate was 77.7%. This means that the system is able to detect almost all of the breath events, at the cost of making 20%+ false detections, mistaking non-breath events for breath events. The high recall rate indicates that the system is successful in identifying key features in actual breath events.

Looking into cases of false detections, most of them were due to exhalation sounds occurring near the end of a phrase [1], where certain singers mix such sounds in their singing. The result itself is understandable because inhalation and exhalation sounds have similar spectral properties, which also means that they are difficult to discriminate. This can be dealt with, for example, by looking into the larger musical context of whether the music/vocal part is actually continuing at that point. The location of the exhalation sounds themselves convey valuable information about the music, so should be processed as such in their own right.

Another case of false detection, though more rare, is the confusion of certain consonants (mostly /h/). As can be expected, /h/ has a spectral envelope similar to that of breath sounds, and so are difficult to discriminate.

As a more exceptional case, the low precision rates of No.011 and No.100 were in part due to background accompaniment sound that crept in from the singer's microphone. This is suggestive of problems to be encountered when dealing with recordings including accompaniment music.

In spite of such problems, the overall high value of R suggests that the detection results of our system has achieved a status of practical value, to be used a preprocessing stage for breath detection. By adding processes to eliminate false detections, our method would become a practial and valuable tool for music processing.

The current system does not make direct use of the 1.6kHz/1.7kHz peak of breath sounds found in the analysis. Incorporating such features is also an important direction for future work.

[1] The total number of the exhalation sounds detected as breath is 175 (335 for all false detections, and 1509 for all events detected as breath).

IV. CONCLUSION

This paper presented a dual approach of analysis/detection for the study of breath sounds in singing. In the analysis phase, novel features of breath sounds were discovered, suggesting importance of further, more detailed research. The results were also demonstrated to be effective for automatic detection, and thus with promising applications.

The current study is based on unaccompanied solo singing, and our future research will extend towards dealing with vocal parts in a more full ensemble performance.

ACKNOWLEDGEMENTS

We thank Dr. Takeshi Saitou (CREST/AIST) and Hiromasa Fujihara (AIST) for their valuable discussions.

REFERENCES

[1] Ruinskiy, D. and Lavner, Y.: "An Effective Algorithm for Automatic Detection and Exact Demarcation of Breath Sounds in Speech and Song Signals", In *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 15, Issue 3, pp.838–850, 2007.

[2] Price, P.J., Ostendorf, M. and Wightman, C.W.: "Prosody and Parsing", In *Proc. DARPA Workshop on Speech and Natural Language*, pp.5–11, 1989.

[3] Wightman, C. W. and Ostendorf, M.: "Automatic Recognition of Prosodic Phrases", In *Proc. ICASSP 91*, pp.321–324, 1991.

[4] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: "RWC Music Database: Popular, Classical, and Jazz Music Databases", in *Proc. 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, 287–288, 2002.

[5] Goto, M. and Nishimura, T.: "AIST Humming Database: Music Database for Singing Research", *The Special Interest Group Notes of IPSJ (MUS)*, 2005(82):7–12, 2005. (in Japanese)

[6] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: "RWC Music Database: Music Genre Database and Musical Instrument Sound Database", in *Proc. 4th International Conference on Music Information Retrieval (ISMIR 2003)*, 229–230, 2003.

[7] Sjolander, K. and Beskow, J.: "WaveSurfer – An Open Source Speech Tool", In *Proc. ICSLP-2000*, Vol.4, pp.464-467, 2000.

[8] Nakano, T., Goto, M. and Hiraga, Y.: "Subjective Evaluation of Common Singing Skills Using the Rank Ordering Method", in *Proc. 9th International Conference of Music Perception and Cognition (ICMPC2006)*, pp.1507–1512, 2006.

[9] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P.: *The HTK Book*, Version 3.2.1, 346p., 2002.