SEVERAL FEATURES FOR DISCRIMINATION BETWEEN VOCAL SOUNDS AND OTHER ENVIRONMENTAL SOUNDS

Shi Yuan-Yuan, Wen Xue, and She Bin

HCI Lab, Samsung Advanced Institute of Technology
phone: +86 10 6842 7711, fax: +86 10 6848 1902, email: yy.shi@samsung.com

ABSTRACT

Several features are found to discriminate between the vocals sounds and other environmental sounds. The vocal sounds include speaking, laughter, etc., 23 kinds of non-verbal and verbal sounds; and the environmental sounds are recorded in domestic environments.

The discriminative features are selected from 22 kinds of features. They are the speech recognition features of LPCC and MFCC, time-spectral features from FFT, statistics of pitch values and contour, ratio of voiced and unvoiced segments, and spectrum of pitch contour. The 9 features calculated from pitch contours perform much better than the features calculated from spectrums, which show no discriminability.

The classification is performed simply by a neural network to evaluate the performance of the 9 features. They are tested on a 21CDs environmental sound database. And the hit rate of 98.73% with the false alarm rate of 11% are obtained. The classification result confirms the effectiveness and efficiency of the features.

1. INTRODUCTION

This paper introduces several discriminative features found for the discrimination between the vocal sounds and other environmental sounds. The work is one part of the environmental sound recognition system and one of its functions is to detect and identify the vocal sounds from other environmental sounds.

The vocal sounds include verbal and non-verbal sounds. Dozens of human sounds are categorized into the "non-verbal sounds", including babble (referring to the babble of children); blow (referring to the nose blow); burp; choke; cough; cry; gasping; groan; hiccup; humming; laughter; scream; sigh; singing; sneeze; snore; throat (referring to the sound of clearing throat); vomit; wheezing; whistle; yawn and yell. And the "verbal sounds" mean the speaking sounds here, but not considering the language meaning of the speech. So there are totally 23 kinds of sounds considered.

Except for the vocal sounds other environmental sounds include 16,000 tracks collected from 21 sound effect CDs, plus the RWCP (Real World Computing Partnership/Real Acoustic Environments Working Group) Sound Scene Database in Real Acoustical Environments [1]. Among them, 5 CDs are human vocal sounds. And 2 CDs are all kinds of non-verbal sounds from persons of different ages and sexes; another 2 CDs are speech, with different ages, sexes, languages, durations and styles; and the left 1 CD is sounds of babies. Except the 5 CDs of vocal sounds left 16 CDs and the RWCP database are all environmental sounds, including more than 10,000 tracks of animal sounds, sounds from human movements and bodies, impact sounds, periodic sounds, sounds in kitchen, sounds in washroom, etc. (A complete list needs pages of tables and is not listed here to save space, and the interesting readers can see the content descriptions of the sound effect CDs launched by Sound Ideas Co. [2])

Almost all kinds of domestic sounds are collected in the database, but except the music. It is not considered currently. Also the vocal sounds and the environmental sounds are all separated into different tracks and not mixed here. So the problem is quite different from and much simpler than the computational auditory scene analysis problem stated by Bregman[3], which must separate and recognize the different sources automatically from one mixed stream.

Section 2 introduces the discriminative features extracted and the performance test for each feature; Section 3 gives the classification test using a neural network; and Section 4 gives the experiment result; finally Section 5 concludes and gives the future work.

2. DISCRIMINATIVE FEATURES

There has not yet been any parameterized model or structured model to describe the transient characteristics of the vocal sounds. But it is believed that the vocal sounds can be detected and identified as for that human can recognize it easily no matter it is a language he never hears before or from a person he never knows. So the first step attracting us is to find out the most discriminative features or cues.

2.1 Feature extraction

Intuitively several features must be considered:

Firstly, the spectral distribution of vocal sounds should be considered. So several features describing the amplitude spectrum envelope are selected as candidates: the speech recognition conventional features of MFCC (Mel-frequency Cepstrum Coefficients) and LPCC (Linear Prediction Cepstrum Coefficients) [4], and FFT spectrum. Also the centroid, bandwidth, roll-off-frequency and bandwidth energy ratio of the FFT spectrum are used to decrease the feature dimension.

Secondly, the pitch characteristics of vocal sounds should be used too. It is believed that the pitch contour is one of the most important cues for vocal sound recognition as for that it stems from the variation of vocal cords vibration, stressness of organ tissues and muscles, pressure of lungs and the breath rate. All the features are unique for human, or mammal. So the features describing the pitch contour and its range and variability are selected as candidates: statistics of fundamental frequency, duration of voiced segment, duration ratio of voiced/unvoiced segments, and the FFT spectrum of pitch contour. Also the centroid, bandwidth and roll-off-frequency of the pitch amplitude spectrum are used to decrease the feature dimension.

Several feature computations are listed below.

- LPCC: 12 coefficients, calculated using the Durbin algorithm described in [5].
- MFCC: 12 coefficients, calculated as in [4].
- FFT: coefficients from 16 FFT to 512 FFT.
- Amplitude spectrum centroid (*Cen*):

$$Cen = \frac{\sum_u u|f(u)|^2}{\sum_u |f(u)|^2} \tag{1}$$

Here, $f(u)$ is the short-term FFT spectrum of the hamming-windowed sound signal, and u is the discrete frequency bin index from 0 to ω_0. ω_0 is the signal Nyquist band (sampling rate/2) in radius.

- Amplitude spectrum bandwidth (*Ban*):

$$Ban^2 = \frac{\sum_u (u - Cen)^2 |f(u)|^2}{\sum_u |f(u)|^2} \tag{2}$$

- Amplitude spectrum roll-off-frequency (*Rof*):

$$Rof = \max\left(h \left| \sum_{u=0}^{h} f(u) < TH * \sum_{u=0}^{M} f(u) \right. \right) \tag{3}$$

TH=0.85.

Here, *M* is the maximal frequency bin index of ω_0.

- Bandwidth energy ratio (*Ber*):

$$Ber_i = \frac{\sum_{u=Li}^{Hi} |f(u)|^2}{\sum_{u=0}^{M} |f(u)|^2}, \ i=0,1,\ldots N\text{-}1 \tag{4}$$

Here the Nyquist band is divided into *N* sub-bands: [0, $\omega_0/2^{N-1}$], [$\omega_0/2^{N-1}, \omega_0/2^{N-2}$], …, [$\omega_0/2, \omega_0$], and $Hi = \omega_0/2^i$, $Li = \omega_0/2^{i+1} = Hi/2$, except L_{N-1}, which is zero.

4 bandwidth energy ratios are calculated separately in the 4 sub-bands of the 512 FFT spectrum here, as in [6~8].

The above 7 features are calculated from the short-term frame of 20ms hamming windowed signal, and the frame shift is 10ms.

- F0 (Fundamental Frequency) statistics: the mean and the variance of F0 values. They are calculated for each track. It is found out that setting F0 zero for unvoiced segment enhances the separation of F0 distributions between the vocal sounds and the environmental sounds to a great extent. So F0 is set zero for unvoiced segments during the pitch tracking process.
- Two duration ratios of voiced and unvoiced segments (*Vtr* and *Zkb*):

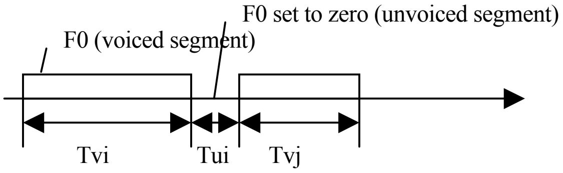

Figure 1. Durations of voiced and unvoiced segments

$$Vtr = \frac{\sum_i Tvi}{\sum_i Tui + \sum_i Tvi} \tag{5}$$

$$Zkb = \frac{Tui}{Tui + Tvi} \tag{6}$$

It is clear from (5) and (6) that *Vtr* is the weighted average of *Zkb*. And *Zkb* is calculated for each pair of voiced and unvoiced segments; and *Vtr* is calculated for each track.

- Spectrum of pitch contour (PitFFT32): The 32 FFT is calculated on every 300ms pitch contour to get the amplitude spectrum of the pitch contour. Also the centroid, bandwidth and roll-of-frequency are calculated.

Finally 22 features are considered. They are LPCC12, MFCC12, FFT16, FFT32, FFT64, FFT128, FFT256, FFT512, CenOfFFT512, BanOfFFT512, RofOfFFT512, Ber4OfFFT512, F0, MeanOfF0, VarOfF0, Vtr, Zkb, PitFFT32, CenOfPitFFT32, BanOfPitFFT32, RofOfPitFFT32 and Zcr (zero-crossing-rate).

2.2 Discrimination performance of each feature

The discrimination performance of each feature depends on the separation degree of the feature distributions between the vocal sounds and the environmental sounds. So the GMM (Gaussian Mixture Model) [4] is used to model the distribution of each feature. Here 50 mixtures are selected experimentally through a series of preliminary tests to find an appropriate number of mixtures. And the diagonal covariance matrix is used for the GMM of the multi-variant feature. Then the GMMs of each feature from vocal sounds are trained through EM algorithm [4].

The discrimination performance of each feature is evaluated by using its Gaussian mixture distribution model of vocal sounds to detect the vocal sound frames from the total sound frames in the database. Different hit rates and false alarm rates can be obtained under the different likelihood threshold conditions. Then the performances of different features can be compared through the detection performance curves of hit rates and false alarm rates.

2.2.1 Test data

The training and testing data are selected from the environmental sound database, including 3 hours of vocal sounds, as listed in Table 1, and 14 hours of environmental sounds, including the 12,000 tracks from 16 CDs and the RWCP database.

1 hour of vocal sounds, including 0.5 hour of non-verbal sounds and 0.5 hour of speech sounds, is used to train the GMM of each feature. Then the left 2 hours of vocal sounds

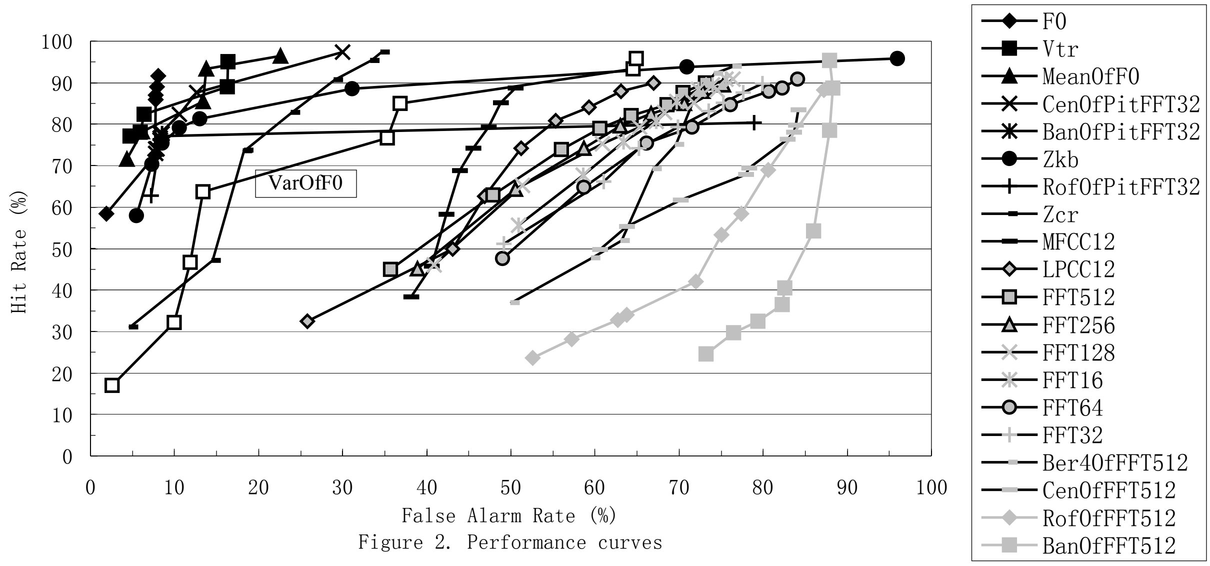

Figure 2. Performance curves

and the total 14 hours of environmental sounds are used in the test.

Table 1. Contents of vocal sounds

Non-verbal	Verbal (speaking)		
Sex	Contents	14 Languages	Sex
Male, female, baby	Broadcasting of news, weather, sports, etc; conversation; movie scene; etc.	English, French, Spanish, Italian, Japanese, Germany, Russian, etc.	Male, female, baby
1.5 Hrs	1.5 Hrs		

2.2.2 Result

Among the 22 features tested, 8 features based on pitch contours, plus Zcr, are discriminative, but others related to the spectrum distributions all fail to detect the vocal sounds. The performance curves are plotted in Figure 2. The legends are ordered according to the performances. It is very clear that the curves in the left-upper region are mostly from the features calculated from the pitch contours; but the ones in the right-lower region are all the features related to the spectrum. So the 9 features from F0 to VarOfF0 are most discriminative and can detect the vocal sounds much better than the other features.

The result is quite different from those using spectrum features to distinguish speech from noise [9~11]. The main reason is that firstly only speech is considered because they are used in a robust speech recognizer or a channel switcher; secondly the noise background discussed is mainly the communication channel noise. So the vocal sounds and the environmental sounds here are more general than previous, and the serious overlap between their feature distributions in the spectrum feature space arises.

So the 9 features of F0, MeanOfF0, VarOfF0, Vtr, Zkb, CenOfPitFFT32, BanOfPitFFT32, RofOfPitFFT32 and Zcr are selected.

3. NEURAL NETWORK CLASSIFICATION

Section 2 only gives the detection performance of each single feature by means of GMM feature distribution likelihood modelling. It is very direct to train a single multivariant GMM for the distribution of the vector composed of the 9 features. Then the detection performance of the 9 features can be evaluated and should be much better than any of the single one.

Also a feature transformation matrix can be applied on the 9 features to remove the linear dependence of different features by means of PCA (Principle Component Analysis) or LDA (Linear Discriminant Analysis) [12]. But it is not critical here because the feature dimension has been very small, and the non-linear classifiers are not sensitive to the dependency of feature dimensions.

Here we prefer to the discriminative non-linear classifier rather than the likelihood classifier of GMM because there are thousands of environmental sounds in the database that can be used to train the classifier discriminatively and enhance the recognition performance. So a neural network is used to test the recognition performance of the 9 features.

The Neural network is the forward-feeding network with 4 layers: layer 1 is 9 neurons as input nodes; layer 2 is 5 to 10 neurons as inner decision nodes; layer 3 is 2 to 5 neurons; and layer 4 is 1 neuron as output node, outputting a value between 0 and 1 as the a posterior probability estimation. It is implemented by Matlab6.5® and the training function *train()* using the backward-propagation algorithm is used. During the train process:

- Feature frames from the vocal sounds and from the environmental sounds are alternatively input to the NN;
- The vocal sound frames are repeatedly batch-used during one batch of the environmental sound frames. The number of the latter is around 5 times the number of the former;
- Frame-asynchronous features are combined into one frame vector by means of combing one long-term feature with the short-term features within its duration time.

Table 2. Hit rates (HR) and false alarm rates (FR) of 12 experiments

ID	1	2	3	4	5	6	7	8	9	10	11	12
Train HR (%)	100	99.22	95.34	98.87	99.52	100	99.24	99.15	99.41	100	98.9	99
Train FR (%)	0	4.73	3.46	0.78	0.2	0.12	0.03	2.67	1.95	0.93	0.67	1.32
Test HR (%)★	**13.16**	**98.73**	**96.5**	**97.4**	**98**	**98**	**99.24**	**100**	**100**	**98.6**	**99.8**	**99.86**
Test FR (%)★	**14.39**	**11**	**10.27**	**10.3**	**11.77**	**10.52**	**8.62**	**10**	**9.87**	**7**	**7.85**	**8**
NN neurons*	5, 2	5, 2	5, 2	10, 5	15, 7	20, 10	5, 2	10, 5	20, 10	5, 2	10, 5	20, 10
Training set size**	1%	10%	50%	50%	50%	50%	10%	50%	50%	10%	50%	50%
	1‰	1%	10%	10%	10%	10%	1%	5%	5%	1%	5%	5%
	Gender-independent						Male			Female		

* "m, n" means m neurons in layer 2 and n neurons in layer 3.

** The upper row of "Training set size" is the percentage of training tracks with the total vocal sounds of 3 hours; and the lower row is the percentage of training tracks with the total environmental sounds of 14 hours.

★Testing set size is the left tracks in the database except the training tracks, so if 10% of the gender-independent vocal sounds are used to train the NN then the left 90% of the gender-independent vocal sounds are used to test the hit rates. Also if 1% of the environmental sounds are used to train then the left 99% of the environmental sounds are used to test the false alarm rates.

Then the a posterior probability estimation is output each frame step of 10ms from the trained network and the averaged value is used as the score of the whole track, which compared with the threshold of 0.5.

4. EXPERIMENT RESULT

Training and testing of the neural network are carried out under the different conditions of training set size, testing set size and network parameters. Results are listed in Table 2. It is concluded from the table that:

- The second row (ID 1) shows a complete separation for the training set but no classification for the testing set, because the training set is too small to train the network with generalization.
- Left tests show performance with around 98% hit rate and 10% false alarm rate;
- Around 99% hit rate and 8% false alarm rate can be obtained in the case of gender-dependent;
- 7 inner neurons are enough for the classification;
- The training set of 10% vocal sound samples, which are around 20 minutes, and 1% environmental sound samples, which are around 10 minutes, are enough for the network training.

So the network classifier can distinguish between the vocal sounds and other environmental sounds quite efficiently, which shows the effectiveness and efficiency of the 9 features.

5. CONCLUSION

9 discriminative features are determined for the classification between the vocal sounds and the complex environmental sounds. It is found out that the pitch-based features win, compared with the spectrum-related features. It confirms that the vocal sounds are recognized by the unique pitch contours, at least. That is, the voiced/unvoiced concatenation in the limited range of V/U transfer frequency (related to the syllable rate of 4Hz [6]); the limited range of duration ratios of V/U segments; and the sliding track of pitch with the limited speed.

The future work includes the analysis on the acoustic and perceptual correlates for the vocal sounds, especially on the pitch and the formants; the modelling of the dynamic characteristics of vocal sounds; and furthermore the recognition between the 23 kinds of vocal sounds.

REFERENCES

[1] S.Nakamura, K.Hiyane, F.Asano and T.Endo, "Sound Scene Data Collection in Real Acoustical Environments", *J. Acoust. Soc. Japan (E)*, 20(3), 1999

[2] Sound Ideas, *http://www.sound-ideas.com*

[3] Albert S. Bregman, *Auditory Scene Analysis*, The MIT Press, Cambridge, Massachusetts, 1990.

[4] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon, *Spoken Language Processing*, Prentice Hall, Inc., New Jersey, 2001.

[5] Xing-Jun Yang, Hui-Sheng Chi, etc. *Speech Signal Digital Processing (in Chinese)*, Electronic Industrial Press, Beijing, 1995.

[6] Eric Scheirer, Malcolm Slaney. "Construction And Evaluation Of A Robust Multifeature Speech/music Discriminator" *Proc. ICASSP'97*, 1997.

[7] Vesa Peltonen, *Computational Auditory Scene Recognition*, M.Sc Thesis, Tampere University of Tech., Finland, 2001.

[8] Dongge Li, Ishwar K. Sethi, Nevenka Dimitrova, and Thomas McGee. "Classification Of General Audio Data For Content-based Retrieval", *Pattern Recognition Letters 22(5)*, pp.533-544, 2001.

[9] Satoh Hideki, Nitta Tsuneo, "Speech Detection Apparatus Not Affected By Input energy Or Background Noise Level", *US Patent 5293588*, Mar. 8, 1994.

[10] Kamiya Shin, Ueda Toru, "Method Of Distinguishing Voice From Noise", *US Patent 4920568*, Apr. 24, 1990.

[11] Yasunaga Satoshi, "Speech Detector Capable Of Avoiding An Interruption By Monitoring A Variation Of A Spectrum Of An Input Signal", *US Patent 4688256*, Aug. 18, 1987.

[12] R.O.Duda and P.E.Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1972.