A NOVEL PITCH EXTRACTION BASED ON JOINTLY TRAINED DEEP BLSTM RECURRENT NEURAL NETWORKS WITH BOTTLENECK FEATURES

Bin Liu, Jianhua Tao, Dawei Zhang and Yibin Zheng

National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190

ABSTRACT

Pitch is an important characteristic of speech and is useful for many applications. However, it is still challenging to estimate pitch in strong noise. In this paper, we propose a joint training approach to determinate pitch. First, a Bidirectional Long Short-Term Memory Recurrent Neural Networks (BLSTMRNN) is trained to map the noisy to clean speech features. Second, the pitch estimation is also a BLSTM-RNN model. The feature mapping neural network serves as a noise normalization module aiming at explicitly generating the clean features which are easier to estimate pitch by the following neural network. BLSTM-RNN is trained on sequential frame-level features and capable of learning temporal dynamics. We also propose to take into account bottleneck features for pitch estimation. The experimental results show that the proposed method can obtain accurate pitch estimation and they show good generalization ability to new speakers and noisy conditions. The proposed approach also significantly outperforms other state-of-the-art pitch estimation algorithms.

Index Terms— Pitch estimation, BLSTM-RNN, feature mapping, joint training, bottleneck features

1. INTRODUCTION

Pitch is an important characteristic of speech. A pitch estimation algorithm robust to background interference is critical to many applications, including speaker identification [1] and speech separation [2]. Although pitch estimation has been studied for decades, it is still challenging to estimate pitch from speech in strong noise. The most prominent difficulty is the corruption of the speech harmonic structure, since most of the existing algorithms depend on a clear harmonic structure [3].

Pitch is supra-segmental feature and context information is important for pitch estimation. Therefore, the pitch estimation can be divided into two steps: pitch candidate selection and pitch tracking. Firstly, possible pitches of each frame are selected as candidates. Then a continuous pitch contour is generated by tracking the selected pitch candidates with the temporal continuity constraint. Dynamic programming [4] or hidden Markov models (HMMs) [5] are often adopted for pitch tracking. For pitch candidate selection, recent studies on robust pitch estimation have explored either harmonic structure in the frequency domain [6, 7], periodicity in the time domain [8, 9], or the periodicity of individual frequency subbands in the time-frequency domain [10, 11]. The traditional methods are mostly based on empirical parameters or a priori assumption on the noise. Rule-based pitch candidate selection may lose useful information because it simply ignores non-peak spectral information. Inspired by the success of deep learning [12, 13], some researchers select pitch candidates with deep models. Han and Wang investigate the use of a deep neural network (DNN) and recurrent neural network (RNN) for pitch candidate selection [14]. Su and Zhang propose to use convolution neural network (CNN) [15]. The deep learning approaches indeed could significantly improve the pitch estimation performance compared with other classification models in the matched noise conditions. But the generalization capability problem to unseen noise conditions was not solved very well.

Motivated by the recent work for noise robust speech recognition and voice activity detection [16, 17, 18], we present a novel feature mapping front-end by using a BLSTM-RNN as a noise normalization module to estimate the clean speech features which make the pitch estimation decision easier with the subsequent BLSTM-RNN. Furthermore, the feature mapping neural network can be jointly trained with the pitch estimation neural network. In addition, we also propose to take into account phonetic information which is represented with bottleneck features for pitch estimation. The use of bottleneck activations from a DNN trained to predict senone posteriors has been previously proposed for the speaker verification [19, 20], language recognition tasks [21, 22] and voice activity detection [23]. In speech synthesis tasks, the phonetic information is significant for pitch tracking [24]. We hypothesize that bottleneck features should also be useful for the pitch estimation task. Therefore, a bottleneck DNN is trained to predict senone posteriors. The activations at the bottleneck layer are then used as input into another model to predict pitch.

There are three main contributions in this paper: (1) a novel pitch extraction based on jointly trained neural network is proposed; (2) BLSTM-RNN is investigated for pitch estimation; (3) the bottleneck features are considered in pitch estimation task.

2. PROPOSED METHOD

In this section, we firstly introduce the framework of the proposed method. Subsequently, the further detail is presented. The flowchart of proposed algorithm is shown in Fig. 1.

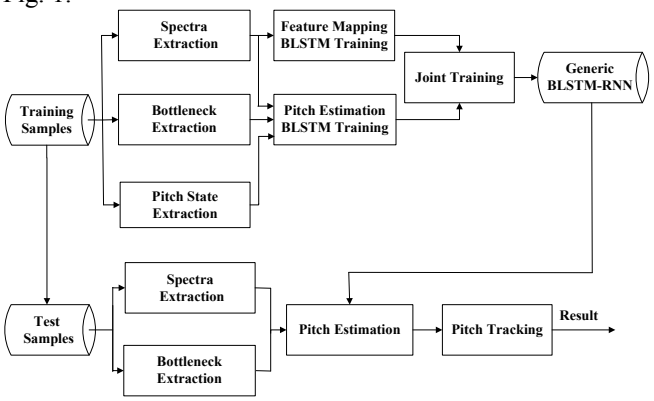

Fig. 1: Block diagram of the proposed method

The overall flowchart of proposed method is illustrated in Fig. 1. In the training stage, the acoustic features of both clean speech and noisy speech training data are extracted. Then two BLSTM-RNN, namely feature mapping BLSTM-RNN and the pitch estimation BLSTM-RNN, are trained. The bottleneck features is considered in pitch estimation BLSTM-RNN. Finally a generic BLSTM-RNN can be generated by joint training of both feature mapping and pitch estimation BLSTM-RNN. In the pitch extraction stage, after the feature extraction, frame-level decision is first given by the generic BLSTM-RNN. To achieve better performance, a post-processing based on dynamic programming can be applied to pitch extraction.

2.1. Spectral features extraction

The features used in this study are extracted from the spectral domain based on [25]. A signal is first decomposed to the spectral domain using short time Fourier transformation. Let $X_t(f)$ denote the power spectral density (PSD) of the frame *t* in the frequency bin *f*. The PSD in the log-frequency domain can be represented as $X_t(q)$, where $q = \log f$. Then, the normalized PSD can be computed as:

$$X_t'(q) = X_t(q)\frac{L(q)}{\bar{X}_t(q)} \tag{1}$$

where *L(q)* represents the long-term average speech spectrum, and $\bar{X}_t(q)$ denotes the smoothed averaged spectrum of speech, which is calculated by using a 21-point moving average filter in the log-frequency domain and averaging over the entire sentence in the time domain in this study. With the normalized spectrum, we further enhance harmonicity using a filter with broadened peaks having an impulse response defined as:

$$h(q) = \begin{cases} \dfrac{1}{\gamma - \cos(2\pi e^q)} - \beta & \text{if } \log(0.5) < q < \log(K + 0.5) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where β is chosen so that $\int h(q)dq = 0$, K indicates the number of harmonics captured by the filter and γ controls the peak width which is set to 1.8.

The convolution $\tilde{X}_t(q) = X_t'(q) * h(q)$ contains peaks corresponding to harmonics and their multiples and submultiples. Only the spectral components in the plausible pitch frequency range (60 to 400 Hz) are selected as features. So we have a spectral feature vector in frame:

$$\tilde{x}_t = (\tilde{X}_t(q_1), ..., \tilde{X}_t(q_n))^T \tag{3}$$

2.2 Pitch features quantization

We set the target pitch frequency from 60 to 400 Hz, a typical range that covers both male and female speech in daily conversations. To simplify the modeling task, we quantize the plausible pitch frequency into pitch states by using 24 bins per octave in a logarithmic scale using [14]. We also incorporate a non-pitched state corresponding to an unvoiced frame. Therefore, we have 68 pitch states: 1 state for the non-pitched frame and the other 67 states for the pitched frame. The output of the model is the probability on pitch states, where each pitch state corresponds to a range of pitch values.

2.3 Bottleneck features extraction

The features generated by the acoustic model are given by the activations in a bottleneck layer of a DNN trained to predict senone posteriors [23]. Senones are defined as tied states within context-dependent phones and are the unit for which observation probabilities are computed during automatic speech recognition (ASR). We can see these bottleneck (BN) features as a low-dimensional representation of the phonetic content in each frame. The spectral features and BN features are combined to train pitch estimation BLSTM-RNN.

2.4. Model training for pitch estimation

The BLSTM-RNN can model the deep representation of long-span acoustic features for pitch estimation. A BLSTM layer consists of a number of recurrently connected such memory blocks which could solve gradient vanish and gradient expansion problem. Each block contains the connected memory cells and three multiplicative units. The surrounding network can only interact with the memory cells via the gates. Two separate recurrent hidden layers are operating in opposite directions, thus providing access to long-range context in both input directions. BLSTM-RNN can be established by stacking multiple RNN hidden layers on top of each other and transform the input sequence. The Back-propagation through time (BPTT) algorithm is applied to both forward hidden nodes and backward hidden nodes, and back-propagates layer by layer [24]. The weight gradient is computed over the entire utterance. The effective learning capability of BLSTM-RNN is expected to benefit pitch estimation. Deep-layered architectures can represent high level representation of input features and BLSTM-RNN can capture information from anywhere in the feature sequence.

The joint training procedure of two BLSTM-RNN can be divided into two steps. The first step is to convert the pitch estimation BLSTM-RNN with the input of noisy spectral features and bottleneck features to the BLSTM-RNN with the input of estimated clean spectral features and bottleneck features, which is implemented via a simple fine-tuning of the original noisy BLSTM-RNN by only changing the input to the estimated clean features and bottleneck features rather than the noisy features and bottleneck features. The second step is to concatenate two BLSTM-RNN to a single generic BLSTM-RNN. We directly stack the pitch estimation layers on top of the feature mapping layers. The output layer of feature mapping and the input layer of pitch estimation are merged in the generic BLSTM-RNN. Using the same objective function as the pitch estimation BLSTM-RNN, all weight and bias parameters are then re-trained. After joint training, the generic BLSTM-RNN yields a better performance than two separated BLSTM-RNN which can be explained as the feature mapping network is refined to enable a better pitch estimation performance rather than optimizing the original MMSE criterion. The model structure of proposed method is shown in Fig. 2.

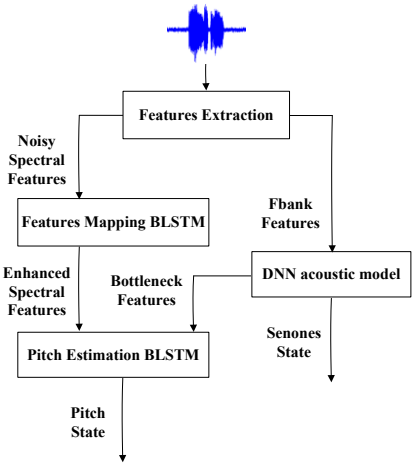

Fig. 2: Model structure of proposed method

2.5. Pitch tracking

Pitch tracking generates a continuous pitch contour by maximizing the pitch probability under the temporal continuity constraint of speech. As suggested in [10], it can be modeled by a Laplacian distribution:

$$p_t(\Delta) = \frac{1}{2\sigma}\exp(-\frac{|\Delta-\mu|}{\sigma}) \quad (4)$$

We generate the final continuous pitch contour by maximizing both the pitch probability and the transfer probability. This process is implemented by a dynamic programming algorithm [15]. The outputs are a sequence of pitch states for a sentence. We convert the sequence of pitch states to the sequence of frequencies and then use a 3-point moving average for smoothing to generate final pitch contours.

3. EXPERIMENT AND RESULTS ANALYSIS

3.1. Data and analysis methodology

We evaluate the performance for the proposed approach using the TIMIT database [26] and GRID database [27]. The sampling rate is 8kHz. The 5000 utterances selected randomly from the TIMIT database were used for training and another 1000 randomly selected utterances from the TIMIT database were used to optimum model parameter. We compared our proposed algorithm with different baseline algorithms on the GRID corpus, where we used the test speakers from No.1 to No.20. The noises used in the training phase include 100 different noise types could be download from [28]. The noise types used in the test set include the training noise types and fourteen new noise types selected from NOISEX-92 [29]. Each utterance is mixed with every noise type in six SNR levels: -5, 0, 5, 10, 15 and 20 dB. The groundtruth pitch is extracted from clean speech using Praat [6]. The same training set and validation set is selected for acoustic model training. The frame-level reference labels of each noisy utterance were generated by forced alignment on the corresponding clean utterance using the acoustic model trained on clean speech data. Sigmoid activation function was used and the number of units in each hidden layer was set to 2048 by default. The bottleneck layer is set to 100, 200 and 400 respectively. The distribution of hidden layer from input to output is 2048-2048-2048-BN-2048.

We evaluate pitch tracking results in terms of two measurements: detection rate (DR) and voicing decision error (VDE) [30]. DR is evaluated on voiced frames, where a F0 estimate is considered correct if the deviation of the estimated F0 is within 5% of the ground truth F0, and VDE indicates the percentage of frames are misclassified in terms of voicing:

$$DR = \frac{N_{0.05}}{N_p} \times 100\% \quad (5)$$

$$VDE = \frac{N_{V\to U} + N_{U\to V}}{N} \times 100\% \quad (6)$$

We compare our approaches with different pitch determination algorithms: PEFAC [25] and a DNN method [14]. Table 1 shows different BLSTM-RNN model.

Table 1. Configure for different BLSTM-RNN model

Configure	Feature Mapping	Jointly Trained	Bottleneck Features
BLSTM-RNN1	No	No	No
BLSTM-RNN2	Yes	No	No
BLSTM-RNN3	Yes	Yes	No
BLSTM-RNN4	Yes	Yes	Yes

3.2. The evaluation of hyper parameter configure

We have conducted experiments for BLSTM-RNN using different numbers of hidden layers and different numbers of hidden units. We firstly evaluate the features enhancement BLSTM-RNN, the BLSTM-RNN with two

hidden layers produces better slightly performance to that with three hidden layers, and outperforms that with one hidden layer obviously. The optimum number of hidden units is 2048. Then we evaluate pitch estimation BLSTM-RNN, the optimum configuration is shown as following: 2 hidden layers and 1024 hidden units (The bottleneck features is 200). In addition, different bottleneck layers (100, 200 and 400) are validated in pitch estimation BLSTM-RNN. The optimum performance is obtained with 200 dimensions features. Finally, we evaluated learning rates in generic BLSTM-RNN. The learning rate is set to 0.0005.

3.3. Overall evaluation

Table 2 shows the detection rates for training noises across a wide range of SNRs. The BLSTM-RNN based methods achieve substantially higher detection rates than others, especially in very low SNR conditions. The jointly trained BLSTM-RNN performs better than the directly trained BLSTM-RNN. Table 3 shows the detection rates for new noises that are not seen in the training phase. Similar to Table 2, the proposed approach yields the best performance in these noise conditions.

Table 2. Performance in trained noise condition for DR

	DR %					
SNR	-5dB	0dB	5dB	10dB	15dB	20dB
PEFAC	42.42	49.43	61.60	72.74	83.94	90.92
DNN	47.34	53.72	64.77	75.06	85.81	91.36
BLSTM-RNN1	49.20	55.33	65.52	75.53	86.06	91.19
BLSTM-RNN2	51.51	57.45	67.33	76.82	86.88	91.46
BLSTM-RNN3	52.02	57.83	67.90	77.46	87.39	91.64
BLSTM-RNN4	**52.54**	**58.28**	**68.33**	**77.82**	**87.64**	**91.71**

Table 3. Performance in unseen noise condition for DR

	DR %					
SNR	-5dB	0dB	5dB	10dB	15dB	20dB
PEFAC	41.37	48.56	60.58	72.42	85.52	91.36
DNN	43.52	50.26	61.63	73.06	85.81	91.48
BLSTM-RNN1	45.65	52.03	61.98	73.14	85.69	90.96
BLSTM-RNN2	47.98	54.32	63.86	74.37	86.32	91.33
BLSTM-RNN3	48.51	54.76	64.37	74.83	86.78	91.49
BLSTM-RNN4	**49.06**	**55.12**	**64.75**	**75.14**	**86.96**	**91.57**

Table 4. Performance in trained noise condition for VDE

	VDE %					
SNR	-5dB	0dB	5dB	10dB	15dB	20dB
PEFAC	35.52	27.26	22.32	16.55	11.37	6.15
DNN	32.11	24.44	19.86	15.07	9.93	5.34
BLSTM-RNN1	30.33	23.02	18.73	14.26	9.34	4.83
BLSTM-RNN2	28.86	21.64	17.63	13.35	8.66	4.57
BLSTM-RNN3	27.66	20.67	16.78	12.69	8.17	4.23
BLSTM-RNN4	**27.03**	**20.19**	**16.38**	**12.34**	**8.02**	**4.15**

Table 5. Performance in unseen noise condition for VDE

	VDE %					
SNR	-5dB	0dB	5dB	10dB	15dB	20dB
PEFAC	36.65	28.07	24.03	18.54	12.73	6.67
DNN	34.04	26.61	22.72	17.58	12.32	6.56
BLSTM-RNN1	32.41	25.63	21.99	17.02	11.89	6.31
BLSTM-RNN2	31.24	24.50	20.73	15.96	11.18	5.69
BLSTM-RNN3	30.04	23.52	19.91	15.38	10.79	5.43
BLSTM-RNN4	**29.47**	**23.05**	**19.52**	**14.93**	**10.46**	**5.22**

Table 4 and Table 5 show the VDE results for the seen and unseen noises, respectively. As shown in the tables, our algorithms produce lower voicing detection errors than others. In addition, the performance could be improved while the bottleneck features are considered in generic BLSTM-RNN.

It was obvious that generic BLSTM-RNN achieved consistent and significant improvements for all the unseen noises with different SNRs, especially at low SNRs. Overall, generic BLSTM-RNN improved the generalization capability, which could be explained as the feature mapping neural network serves as a noise normalization module aiming at explicitly generating the clean features which are easier to estimate pitch by the following neural network. The gap between proposed BLSTM-RNN and others BLSTM-RNN became larger at lower SNR for the same noise type, which demonstrated proposed method was more effective under low SNRs. In addition, the bottleneck features are useful for the pitch estimation, because they are trained as a low-dimensional representation of the phonetic content in each frame and the phonetic information is significant for pitch tracking.

4. CONCLUSION

In this paper, we propose a joint training approach to determinate pitch. BLSTM-RNN is investigated for pitch estimation. We also propose to take into account phonetic information which is represented with bottleneck features for pitch estimation. The experimental results show the proposed approach also significantly outperforms other state-of-the-art pitch estimation algorithms.

As for the future work, we will focus on improving the practicability of our approach in both accuracy and efficiency. We plan to explore different structures for the bottleneck DNN. We will extend our method to multi-pitch estimation.

5. ACKNOWLEDGEMENTS

This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No.2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61425017, No.61403386, No. 61305003), the Strategic Priority Research Program of the CAS (GrantXDB02080006) and the Major Program for the National Social Science Fund of China (13&ZD189).

6. REFERENCE

[1] Atal B S. Automatic speaker recognition based on pitch contours[J]. The Journal of the Acoustical Society of America, 1972, 52(6B): 1687-1697.

[2] Han K, Wang D L. A classification based approach to speech segregation[J]. The Journal of the Acoustical Society of America, 2012, 132(5): 3475-3483.

[3] Zhengwei Huang. Multi-pitch estimation[C]. In Proceedings of the ACM International Conference on Multimedia, vol. 1, pp. 801–804, 2014.

[4] Chu W, Alwan A. SAFE: a statistical approach to F0 estimation under clean and noisy conditions[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(3): 933-944.

[5] Jin Z, Wang D L. HMM-Based Multipitch Tracking for Noisy and Reverberant Speech.[J]. IEEE Transactions on Audio Speech & Language Processing, 2011, 19(5):1091-1102.

[6] Boersma P, Weenink D. Praat: doing phonetics by computer (version 5.3. 41)[J]. Computer program. Retrieved, 2013, 1.

[7] Nakatani T, Amano S, Irino T, et al. A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments[J]. Speech Communication, 2008, 50(3): 203-214.

[8] Talkin D. A robust algorithm for pitch tracking (RAPT)[J]. Speech coding and synthesis, 1995, 495: 518.

[9] De Cheveigné A, Kawahara H. YIN, a fundamental frequency estimator for speech and music[J]. The Journal of the Acoustical Society of America, 2002, 111(4): 1917-1930.

[10] Wu M, Wang D L, Brown G J. A multipitch tracking algorithm for noisy speech[J]. IEEE Transactions on Speech and Audio Processing, 2003, 11(3): 229-241.

[11] Huang F, Lee T. Pitch estimation in noisy speech using accumulated peak spectrum and sparse estimation technique[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(1): 99-109.

[12] Ciregan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification[C]//Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012: 3642-3649.

[13] Huang Z, Dong M, Mao Q, et al. Speech emotion recognition using CNN[C]//Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 801-804.

[14] Han K, Wang D L. Neural networks for supervised pitch tracking in noise[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014: 1488-1492.

[15] Su H, Zhang H, Zhang X, et al. Convolutional neural network for robust pitch determination[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 579-583.

[16] Gao T, Du J, Dai L R, et al. Joint training of front-end and back-end deep neural networks for robust speech recognition[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015: 4375-4379.

[17] Narayanan A, Wang D L. Joint noise adaptive training for robust automatic speech recognition[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014: 2504-2508.

[18] Wang Q, Du J, Bao X, et al. A universal VAD based on jointly trained deep neural networks[C]//Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[19] McLaren M, Lei Y, Ferrer L. Advances in deep neural network approaches to speaker recognition[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015: 4814-4818.

[20] Richardson F, Reynolds D, Dehak N. A unified deep neural network for speaker and language recognition[J]. arXiv preprint arXiv:1504.00923, 2015.

[21] Song Y, Jiang B, Bao Y B, et al. I-vector representation based on bottleneck features for language identification[J]. Electronics Letters, 2013, 49(24): 1569-1570.

[22] Matejka P, Zhang L, Ng T, et al. Neural network bottleneck features for language identification[J]. Proc. of IEEE Odyssey, 2014: 299-304.

[23] Ferrer L, Graciarena M, Mitra V. A phonetically aware system for speech activity detection[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 5710-5714.

[24] Fan Y, Qian Y, Xie F L, et al. TTS synthesis with bidirectional LSTM based recurrent neural networks[C]//Interspeech. 2014: 1964-1968.

[25] Gonzalez S, Brookes M. PEFAC-a pitch estimation algorithm robust to high levels of noise[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(2): 518-530.

[26] Garofolo J S. Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database[J]. National Institute of Standards and Technology (NIST), Gaithersburgh, MD, 1988, 107.

[27] Cooke M, Barker J, Cunningham S, et al. An audio-visual corpus for speech perception and automatic speech recognition[J]. The Journal of the Acoustical Society of America, 2006, 120(5): 2421-2424.

[28] G. Hu, 100 Nonspeech Sounds 2006 [Online]. Available: http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html

[29] Varga A, Steeneken H J M. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems[J]. Speech communication, 1993, 12(3): 247-251.

[30] Sun X. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio[C]//Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on. IEEE, 2002, 1: I-333-I-336.