MULTI-SPEAKER PITCH TRACKING VIA EMBODIED SELF-SUPERVISED LEARNING

Xiang Li, Yifan Sun, Xihong Wu, Jing Chen

Department of Machine Intelligence, Speech and Hearing Research Center,
and Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China

ABSTRACT

Pitch is a critical cue in human speech perception. Although the technology of tracking pitch in single-talker speech succeeds in many applications, it's still a challenging problem to extract pitch information from mixtures. Inspired by the *motor theory* of speech perception, a novel multi-speaker pitch tracking approach is proposed in this work, based on an embodied self-supervised learning method (EMSSL-Pitch). The conceptual idea is that speech is produced through an underlying physical process (i.e., human vocal tract) given the articulatory parameters (articulatory-to-acoustic), while speech perception is like the inverse process, aiming at perceiving the intended articulatory gestures of the speaker from acoustic signals (acoustic-to-articulatory). Pitch value is part of the articulatory parameters, corresponding to the vibration frequency of vocal folders. The acoustic-to-articulatory inversion is modeled in a self-supervised manner to learn an inference network by iteratively sampling and training. The learned representations from this inference network can have explicit physical meanings, i.e., articulatory parameters where pitch information can be further extracted. Experiments on GRID database show that EMSSL-Pitch can achieve a reachable performance compared with supervised baselines and be generalized to unseen speakers.

Index Terms— Multi-pitch tracking, self-supervised learning, speech perception, speech production

1. INTRODUCTION

Conventional pitch tracking algorithms have succeeded in the single-speaker scenario [1, 2, 3, 4], but failed to track pitch contours from multi-speaker mixtures. Such so-called multi-pitch tracking has been a long-standing challenge during the last couple decades. The recent introduction of deep learning has dramatically accelerated the progress and boosted the performance of this task. Existing approaches can be conceptually regarded as learning deep neural networks to estimate the frame-level probabilities of pitch periods, followed by the factorial hidden Markov model (FHMM) for continuous pitch tracking [5, 6, 7, 8, 9, 10].

Although different training frameworks are proposed, those pitch tracking approaches are essentially performed in a supervised learning manner where the generalization is the most common issue. Moreover, the requirement that the reference pitch should be extracted from the single-speaker speech in advance, leads to a more serious generalization problem when the trained model is applied to real data. In principle, pitch is linked to an underlying physical process during speech production. There should be a modeling process with physical meaning, instead of a simple acoustic-to-pitch data mapping.

From the perspective of speech production, when human produces voiced sounds, the glottis releases regular pulses of air into the vocal tract, resulting in periodicity in the acoustic waveform. This periodicity, known as the fundamental frequency (F0), is the principal determinant of perceived pitch [11]. Hence, speech production can be modeled as an articulatory-to-acoustic conversion. When it comes to speech perception, Liberman et al. [12] propose the *motor theory* which indicates the objects of speech perception are the intended phonetic or articulatory gestures of the speaker when producing an utterance. This process is like the acoustic-to-articulatory conversion, an inverse of speech production. As mentioned before, production process corresponds to the human vocalization mechanism, which already can be simulated through a tube model by waveguide techniques, such as Tube Resonance Model (TRM) [13]. Given the articulatory parameters, TRM can synthesize the corresponding speech. Pitch value is part of the input articulatory parameters, corresponding to the vibration frequency of vocal folders. Accordingly, the acoustic-to-articulatory problem is rarely studied, but it is actually the pitch tracking task concerned, in which the articulatory parameters (e.g. F0) are desired to be learned from observed acoustic data.

Recently, an embodied self-supervised learning (EMSSL) method is proposed, to tackle the acoustic-to-articulatory problem [14]. In this work, it is extended to the multi-speaker scenario and pitch tracking task, resulting in EMSSL-

This work was supported by the National Key Research and Development Program of China (Grant No.2021ZD0201503), a National Natural Science Foundation of China (Grant No. 12074012), a research funding from SONOVA, and High-performance Computing Platform of Peking University.

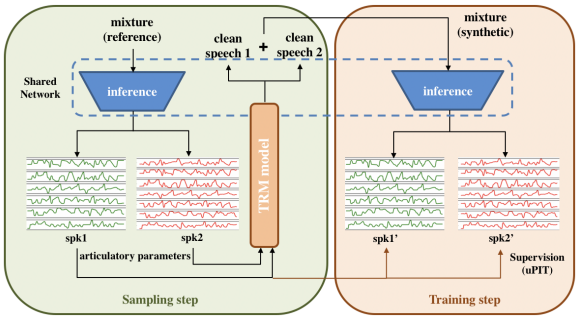


Fig. 1: EMSSL-Pitch overview for one iteration.

Pitch, a novel multi-pitch tracking approach inspired by the *motor-theory* of speech perception and performed in a self-supervised manner. EMSSL works in an analysis-by-synthesis manner to learn an inference network by iteratively sampling and training. The entire learning procedure integrates the physical generation process to obtain meaningful parameters from acoustic observations in a self-supervised manner. When it is extended to our case, EMSSL-Pitch aims to simultaneously estimate articulatory parameters for each individual from observed mixtures. Pitch information can be further detached to perform multi-pitch tracking. The most prominent characteristic of EMSSL-Pitch is that it is the first self-supervised approach for multi-pitch tracking task, and it is possible to be adapted to unseen speakers. Additionally, the inference network is physically meaningful, corresponding to an inversion of simulated vocal tract model. EMSSL-Pitch is evaluated through comparison with supervised baselines and in terms of generalization to unseen speakers.

2. EMSSL-PITCH FRAMEWORK

2.1. Overview

The purpose of EMSSL-Pitch is to learn an acoustic-to-articulatory inversion, which can generate articulatory parameters for each speaker from observed acoustic mixtures. We adopt the existing EMSSL method and extend it to the multi-speaker pitch tracking.

In EMSSL, this inverse problem is tackled by learning an inference network while employing the corresponding physical forward process. A novel analysis-by-synthesis procedure is proposed by iteratively sampling and training. When it is extended to EMSSL-Pitch as shown in Figure 1, at the sampling step, given an observed 2-speaker mixture, the inference network is aimed to approximate the articulatory parameters for each speaker which are then fed to a physical process separately. The synthetic speech for each speaker are finally summed; At the training step, the same network is trained on the generated paired data to predict articulatory parameters from observations. These two steps operate iteratively and boost each other, similar to the guess-try-feedback process in human learning. Pitch contours can be extracted from the learnt articulatory parameters for each speaker, to perform multi-pitch tracking task.

Algorithm 1 Embodied Self-supervised Learning

Input: Observation dataset $\{\mathbf{x}_i\}_{i=1}^N$, traininng set $\Phi = \emptyset$, forward operator F, inference neural network R_θ, sample number per datapoint L, iteration limit T, epoch number E, batch size M
Output: optimized parameters θ

$\theta \leftarrow$ Random initialize parameters
for iteration t = 1 to T **do**
  *** sampling step start ***
  $\theta^t \leftarrow \theta$ (Set the latest θ for calculating posterior in iteration t)
  $\Phi_t \leftarrow \emptyset$ (Initialize an empty set of sampled paired data for current iteration)
  for each $\mathbf{x}_i$ in $\{\mathbf{x}_i\}_{i=1}^N$ **do**
    $Q_{t,i}(\mathbf{z}) \leftarrow P(\mathbf{z}|R_{\theta^t}(\mathbf{x}_i))$ (Approximate posterior)
    $\{\mathbf{z}_{t,i}^{(l)}\}_{l=1}^L \leftarrow$ Draw L samples from $Q_{t,i}(\mathbf{z})$
    $\{(\mathbf{z}_{t,i}^{(l)}, \mathbf{x}_{t,i}^{(l)})\}_{l=1}^L \leftarrow$ Generate paired data, $\mathbf{x}_{t,i}^{(l)} = F(\mathbf{z}_{t,i}^{(l)})$
    $\Phi_t \leftarrow \Phi_t \bigcup \{(\mathbf{z}_{t,i}^{(l)}, \mathbf{x}_{t,i}^{(l)})\}_{l=1}^L$(Update current sampling set)
  $\Phi \leftarrow$ Update training set Φ with newly sampled paired data set Φ_t
  *** training step start ***
  for epoch e = 1 to E **do**
    for number of batches in a training epoch **do**
      $\{(\mathbf{z}^{(m)}, \mathbf{x}^{(m)})\}_{m=1}^M \leftarrow$ Random minibatch of M samples from training set Φ
      $\theta \leftarrow \nabla_\theta \frac{1}{M} \sum_{m=1}^M -logP(\mathbf{z}^{(m)}|R_\theta(\mathbf{x}^{(m)}))$ (Update parameters using gradients)
return θ

2.2. Embodies Self-supervised Learning (EMSSL)

EMSSL is proposed originally to solve the inverse problem of the following forward process: $\mathbf{x} = F(\mathbf{z}) + \mathbf{e}$, where $\mathbf{x} \in X$ is the observation and $\mathbf{z}$ is the input parameter to the forward operator F; $\mathbf{e} \in X$ is the observational noise and included into F in the following description for clarity. In our case, the objective is to learn an inverse of F, in order to infer parameters from acoustic observations. Specifically, given $Ni.i.d.$ observed data $\{\mathbf{x}_i\}_{i=1}^N \subset X$ and a forward operator F, EMSSL aims to find an inverse operator $R_{\hat{\theta}}$: $X \rightarrow Z$:

$$\hat{\theta} := \underset{\theta}{argmin} \frac{1}{N} \sum_{i=1}^{N} -logP(\mathbf{x}_i|R_\theta(\mathbf{x}_i)) \tag{1}$$

where $-logP(\mathbf{x}|R_\theta(\mathbf{x}))$ is the reconstruction error. If observational noise is additive Gaussian, then $-logP(\mathbf{x}|R_\theta(\mathbf{x})) \propto \left\|\mathbf{x} - F(R_\theta(\mathbf{x}))\right\|^2$.

However, F corresponds to a physical process, i.e., vocal tract model, which is non-linear and not differentiable w.r.t. its parameters. Since it is difficult to apply analytical methods, EMSSL considers to solve the inverse problem through self-supervised learning with iteratively sampling and training. Details are shown in Algorithm 1.

2.3. Extension to Multi-pitch Tracking (EMSSL-Pitch)

In this work, EMSSL is extended to the two-speaker scenario and pitch tracking task where observation data is two-speaker mixtures and Tube Resonance Model (TRM) [13] is adopted as the forward operator. TRM simulates the propagation of sound waves through a tube by waveguide techniques, served as the articulatory synthesizer.

Given observed two-speaker mixtures and the TRM model, EMSSL-pitch aims to learn an inference model

through iterations, producing the articulatory parameters for each speaker. Each iteration involves a sampling step and a training step in the same way as the original EMSSL. In detail, at sampling step, the inference network with two output branches is used to approximate the articulatory parameters for each speaker (articulatory), which are sampled and fed into the TRM separately to synthesize speech. Synthesized speech of each speaker is finally summed (synthetic mixture). At the training step, the same inference network is trained on the generated paired (synthetic mixture-articulatory) data in a supervised manner. When the learning procedure ends, the inference network can be run in a deterministic way to produce inversion results (i.e., articulatory parameters).

Since the inference network will output two sets of articulatory parameters, each for one speaker in mixtures, label permutation problem should be addressed. Hence, utterance-level permutation invariant training (uPIT) [15] is adopted at the training step. Finally, pitch information can be further extracted from the learned articulatory parameters. Among them, *reference_glottal_pitch* in utterance-rate parameters is regarded as the mean pitch and *microInt* in control-rate parameters is the relative variations. We combine them to obtain the final absolute pitch contours for each speaker.

3. EXPERIMENTAL SETUP

3.1. Articulatory Synthesizer

The Tube Resonance Model (TRM) [13] is adopted as the forward process in the EMSSL-Pitch framework. It simulates the propagation of sound waves through a tube by waveguide techniques, composed of a vocal tract with 8 segments and a nasal cavity with 5 segments. To synthesize speech, the TRM takes two kinds of parameters as input: 26-d utterance-rate parameters and 16-d control-rate parameters. The former describes the global state of the tube and the latter dynamically controls the tube to produce time varying speech by changing diameters of segments and velum. More details of the TRM are provided in supplementary materials of [14].

3.2. Model Settings

For the inference model in EMSSL-Pitch, a temporal convolutional network (TCN) is adopted as its successful application in speech separation [16] to replace conventional RNN. The TCN consists of 8 stacked 1-D dilated convolutional blocks, each with different dilation factors $1, 2, 4, ..., 2^{8-1}$ repeated 3 times. Two linear layers are stacked on top of the TCN, separately, followed by tanh activation, to map the representations of each frame to 13-d utterance-rate parameters and 16-d control-rate parameters, respectively. Mean square errors are used for the loss of both utterance-rate parameters L_u and control-rate parameters L_c. The optimized objective is to minimize $L_u + \lambda L_c$, where λ is a super-parameter. uPIT is applied during the training process.

3.3. Training Details

We conduct experiments on the GRID database [17], which consists of 1,000 sentences from each of 34 speakers. Mixtures are generated by following the same way as the baseline system [10]. We use 80-dimensional log magnitude mel-spectrogram with 50 ms frame length and 12.5 ms frame shift.

The neural network is randomly initialized and trained with the loss weight $\lambda = 0.001$. We use the Adam optimizer with learning rate of 5×10^4, with $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and train the model on a computer with 2 NVIDIA V100 GPUs and 2 Xeon E5-2690 V4 CPUs.

3.4. Evaluation Metric

The error measurement E_{Total} [9] is used to evaluate the pitch accuracy. It jointly evaluates the performance in terms of pitch accuracy and speaker assignment, by combining the percentile representation of voicing decision, permutation error, gross error and fine error. The smaller E_{Total} value is, the better performance indicates. Reference pitch (ground-truth) is extracted from pre-mixed single-speaker utterances using the RAPT algorithm [1].

4. RESULTS

4.1. Preliminary on Single Speaker

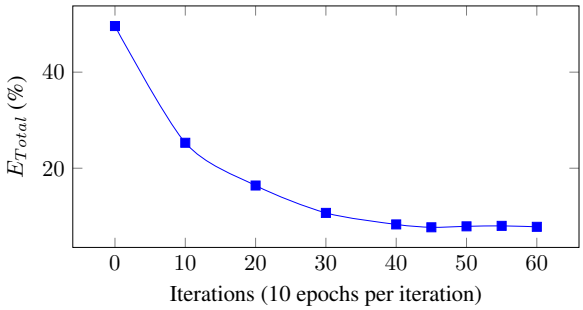


Fig. 2: $E_{Total}(\%)$ of EMSSL on single-speaker tracking.

Before we conduct the formal experiments on two-speaker mixtures, EMSSL is first evaluated on pitch tracking through preliminary experiment with single-speaker speech as the observational input. E_{Total} scores at different iterations are plotted in Figure 2. The component of average pitch deviation in E_{Total} is calculated as: $\nabla f_m = \left|\hat{f}_m - f_m\right| / f_m$, where $\hat{f}_m$ is extracted pitch values on articulatory parameters from inference model and f_m is reference pitch values from PRAT algorithm. Hence, the results can be regarded as the accuracy of exact pitch values from EMSSL, compared with the ground-truth, which shows the potential of such self-supervised learning approach on pitch tracking task. With the number of iteration increasing, E_{Total} degrades rapidly at the beginning and then tends to oscillate stably around 8%.

Table 1: $E_{Total}(\%)$ of of different multi-pitch tracking systems with different gender combinations.

System	Same-Gender				Different-Gender			
	0 dB	3 dB	6 dB	9 dB	0 dB	3 dB	6 dB	9 dB
Wohlmayr et al. SD [10]	31.0	31.5	32.6	34.1	26.0	26.6	27.1	28.3
SPD-Pitch [10]	12.4	12.4	12.8	13.9	11.5	11.6	11.8	12.5
GPD-Pitch [10]	25.7	26.0	27.3	29.6	**14.3**	**14.6**	**15.2**	**16.3**
uPIT-Pitch [10]	**25.1**	**24.9**	**24.4**	**25.2**	14.8	14.8	15.4	16.7
EMSSL-Pitch-v1	26.9	26.1	26.6	27.3	15.8	16.1	16.5	17.3
EMSSL-Pitch-v2	28.3	27.8	28.2	28.9	17.6	18.0	18.4	19.2
EMSSL-Pitch-v2-w/o FT	35.3	34.8	35.2	35.9	30.6	31.4	31.9	32.5
EMSSL-Pitch-v2-FT	28.9	28.3	28.6	29.3	20.9	21.6	22.3	22.8

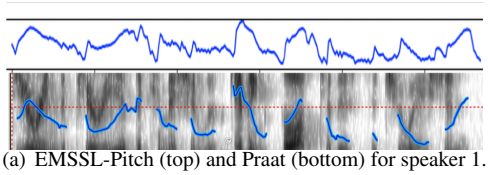
4.2. Evaluation on Two-speaker Mixtures

When the original EMSSL is extended to the multi-speaker scenario, EMSSL-Pitch should be evaluated from two aspects: how is the performance when extended from single- to multi-speaker scenario; how is the performance when compared with other existing systems, which determines different reference pitch should be used during E_{Total} calculation. Therefore, the former aspect applies the extracted pitch from the inference model trained on single-speaker speech as the reference, denoted as EMSSL-Pitch-v1, while the latter applies the extracted ground-truth pitch from RAPT algorithm, denoted as EMSSL-Pitch-v2.

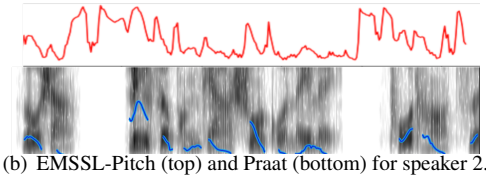
Table 1 shows the E_{Total} results of different multi-pitch tracking approaches with different gender combinations. We take the results list in [10] as the supervised baselines, which can be categorized as: speaker-dependent (SD and SPD-Pitch) and speaker-independent (i.e., GPD- and uPIT-Pitch) approaches. SPD-Pitch shows the best results due to the usage of speaker-dependent information. Within speaker-independent systems, EMSSL-Pitch-v1 can achieve a slightly worse but reachable performance compared with the other two baselines. The performance becomes worse for EMSSL-Pitch-v2 where using the reference pitch from RAPT. It indicates there is still a performance gap between the inference model and ground truth. However, it is noteworthy that the proposed system is self-supervised while the baselines are all based on supervised learning.

4.3. Generalization to Unseen Two-speaker Mixtures

Since the EMSSL is claimed to be performed in self-supervised manner, it should be evaluated in terms of its generalization ability. We conduct the adaptive experiments where inference model is pre-trained on GRID database and then tested on WSJ0-2mix corpus with and without fine-turned (FT). The bottoms two rows in Table 1 show the results when EMSSL-Pitch is adapted with increasing iterations. The final E_{Total} can be around 25% and 33% with and w/o FT, respectively.



(a) EMSSL-Pitch (top) and Praat (bottom) for speaker 1.



(b) EMSSL-Pitch (top) and Praat (bottom) for speaker 2.

Fig. 3: Comparison between model outputs and ground-truth.

To intuitively evaluate the quality of extracted pitch contour from EMSSL-Pitch, we plot the results from the model against those from Praat [18]. In Figure 3, we can figure out some part of the error comes from the voicing decision that assigns pitch values to those unvoiced segments. Therefore, we apply additional control-rate parameter *glotVol* as the voice/unvoice decision and E_{Total} score can be reduced by 1.8% for EMSSL-Pitch-v1 system.

5. CONCLUSION AND DISCUSSION

In this work, a novel multi-pitch tracking approach inspired from *motor theory* in speech perception is proposed. It is extended from the existing embodied self-supervised learning method, denoted as EMSSL-Pitch. EMSSL-Pitch is first applied in the single-speaker scenario to verify its effectiveness on pitch tracking, and then evaluated in terms of two-speaker pitch tracking, which results in a slightly worse performance than supervised baselines. However, it can be generalized to unseen speakers after fine-tuned due to its self-supervised nature. In our future work, information from other dimensions in the articulatory parameters can be integrated to further improve the accuracy of extracted pitch.

6. REFERENCES

[1] David Talkin and W Bastiaan Kleijn, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, pp. 518, 1995.

[2] Alain De Cheveigné and Hideki Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[3] Yuzhou Liu and DeLiang Wang, "Robust pitch tracking in noisy speech using speaker-dependent deep neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5255–5259.

[4] Yuzhou Liu and DeLiang Wang, "Time and frequency domain long short-term memory for noise robust pitch tracking," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5600–5604.

[5] Francis R Bach and Michael I Jordan, "Discriminative training of hidden markov models for multiple pitch tracking [speech processing examples]," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* IEEE, 2005, vol. 5, pp. v–489.

[6] Mads Græsbøll Christensen and Andreas Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech & Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[7] Zhaozhang Jin and DeLiang Wang, "Hmm-based multipitch tracking for noisy and reverberant speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1091–1102, 2010.

[8] Mingyang Wu, DeLiang Wang, and Guy J Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, 2003.

[9] Michael Wohlmayr, Michael Stark, and Franz Pernkopf, "A probabilistic interaction model for multipitch tracking with factorial hidden markov models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 799–810, 2010.

[10] Yuzhou Liu and DeLiang Wang, "Permutation invariant training for speaker-independent multi-pitch tracking," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5594–5598.

[11] John C Middlebrooks, Jonathan Z Simon, Arthur N Popper, and Richard R Fay, *The auditory system at the cocktail party*, vol. 60, Springer, 2017.

[12] Alvin M Liberman and Ignatius G Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, no. 1, pp. 1–36, 1985.

[13] David R Hill, Craig R Taube-Schock, and Leonard Manzara, "Low-level articulatory synthesis: A working text-to-speech solution and a linguistic tool 1," *Canadian Journal of Linguistics/Revue canadienne de linguistique*, vol. 62, no. 3, pp. 371–410, 2017.

[14] Yifan Sun and Xihong Wu, "Embodied self-supervised learning by coordinated sampling and training," *arXiv preprint arXiv:2006.13350*, 2020.

[15] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[16] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[17] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[18] Paul Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.