PGSS: Pitch-Guided Speech Separation

Xiang Li, Yiwen Wang, Yifan Sun, Xihong Wu, Jing Chen

School of Intelligence Science and Technology, Peking University, Beijing, China
chenj@cis.pku.edu.cn

Abstract

Monaural speech separation aims to separate concurrent speakers from a single-microphone mixture recording. Inspired by the effect of pitch priming in auditory scene analysis (ASA) mechanisms, a novel pitch-guided speech separation framework is proposed in this work. The prominent advantage of this framework is that both the permutation problem and the unknown speaker number problem existing in general models can be avoided by using pitch contours as the primary means to guide the target speaker. In addition, adversarial training is applied, instead of a traditional time-frequency mask, to improve the perceptual quality of separated speech. Specifically, the proposed framework can be divided into two phases: pitch extraction and speech separation. The former aims to extract pitch contour candidates for each speaker from the mixture, modeling the bottom-up process in ASA mechanisms. Any pitch contour can be selected as the condition in the second phase to separate the corresponding speaker, where a conditional generative adversarial network (CGAN) is applied. The second phase models the effect of pitch priming in ASA. Experiments on the WSJ0-2mix corpus reveal that the proposed approaches can achieve higher pitch extraction accuracy and better separation performance, compared to the baseline models, and have the potential to be applied to SOTA architectures.

Introduction

The human auditory system shows the extraordinary ability at selectively attending to the target speech in the presence of interference, which is described as the so-called cocktail party problem (Cherry 1953). Bregman attributes this ability to "auditory scene analysis" (ASA) (Bregman 1994), a process of auditory stream formation and segregation. Specifically, the formation stage aims to form a single auditory stream from an incoming mixture, by successive auditory periphery analysis, perceptual features extraction (e.g., pitch, timbre, spatial location) (Middlebrooks et al. 2017) and temporal coherence analysis (Shamma, Elhilali, and Micheyl 2011). In the segregation stage, attention mechanism (Bey and McAdams 2002), as well as prior experience (e.g., familiarity, priming) (Snyder et al. 2008, 2009a,b; Snyder and Weintraub 2011; Riecke et al. 2009, 2011; McClelland, Mirman, and Holt 2006) would serve to enhance the perception of a particular stream (foreground), while suppressing others (background).

In the above ASA process, pitch plays an important role. Pitch, often corresponds to the fundamental frequency (F0) of harmonics. In the psychoacoustics aspect, a series experiments using from alternating pure tones (Shamma, Elhilali, and Micheyl 2011), synthetic vowels (Broadbent and Ladefoged 1957) to natural continuous speech (i.e., multi-speaker scenario) (Glasberg and Moore 1986; Darwin and Hukin 2000; Darwin, Brungart, and Simpson 2003) showed the F0 difference substantially contributed to the perceptual segregation. Other groups of research find that priming listeners with perceptual cues (e.g., pitch) of the target speaker could help them attend to the target speaker when competing speakers are present (Brungart 2001; Freyman, Balakrishnan, and Helfer 2004; Kidd Jr, Mason, and Gallun 2005). Church and Schacter (Church and Schacter 1994) showed there was a significant priming effect when the words were spoken with the same fundamental frequency at study and test, which means presenting listeners with the priming word spoken at the same pitch as the target test word can improve the identification performance. In the aspect of computational models, the most representative work is computational auditory scene analysis (CASA) (Brown and Cooke 1994; Wang and Brown 2008), where pitch as a discriminative cue, can be used to improve the speech separation performance (Hu and Wang 2010, 2013; Wang, Soong, and Xie 2019).

Due to the robustness of the auditory system in the complex acoustic scene, the computational model for the ASA process has been explored a lot, in order to extract target speech from competing inferences, i.e, speech separation. Recently, the development of deep learning brings opportunities to tackle this task. A supervised learning framework was proposed to learn a mapping from mixture to separated speakers through neural networks. Most existing models separate each speaker simultaneously, which suffers from permutation problem (Kolbaek et al. 2017) and output dimension mismatch problem (Hershey et al. 2016). The former is related to how the output layers are tied to the underlying speakers. The latter arises from an unfixed number of speakers in the mixture, leading to unfixed output layers. Although permutation invariant training (PIT) (Yu et al. 2017; Kolbaek et al. 2017) and deep clustering (DC) (Hershey et al. 2016) are proposed successively to address these two problems, respectively, both of them cannot deal with the problem of unknown number of speakers.

In the ASA mechanism, it seems that humanity is not generally capable of attending to every aspect of the auditory input, but primarily to one stream at a time. This attended stream then stands out perceptually, while the rest of the sound is less prominent (Bregman 1994; Middlebrooks et al. 2017). Therefore, during the computational modeling, the more reasonable way is to separate one specific speaker each time instead of all speakers simultaneously. The key problem then comes to how to define this attended stream, i.e., target speaker. According to the effect of pitch priming, pitch can be utilized as the perceptual cue to indicate this target.

To address the problems mentioned before, and more importantly, to integrate the effect of pitch priming in the ASA mechanism, this work proposes a novel pitch-driven speech separation framework. It can be further divided into two phases: pitch extraction and speech separation. The former aims to extract pitch contours for each speaker from mixtures, which can be regarded as modeling the process of bottom-up perceptual cue extraction. In the latter phase, a speech separation model conditioned on the pitch is proposed, based on a conditional generative adversarial network (CGAN) (Mirza and Osindero 2014). Any pitch contour belonging to a specific speaker from the first phase can be selected as the condition, and the model will only output the corresponding speaker. Such a model naturally models the role of pitch prime. Experiments on WSJ0 show the proposed framework can achieve comparable performance with the Conv-TasNet model (Luo and Mesgarani 2019) and meanwhile it overcomes the limitation caused by permutation and unknown speaker number problems (Takahashi et al. 2019).

Contributions: (1) We proposed a novel pitch-driven speech separation framework, where only one speaker is separated each time corresponding to the given pitch condition. Such framework avoids both permutation and unknown speaker number problems, and models the effect of pitch priming in the ASA process; (2) Within conditional separation frameworks, we utilize pitch contour instead of conventional speaker embedding, as the condition. Pitch, as a low-level and short-term perceptual feature, can still be discriminative enough on feature space for those speakers with similar timbre; (3) Instead of the conventional mask-based method, the proposed method produces waveform directly, resulting in improved quality of separated speech.

Related Work

All systems described in this paper operate on monaural recordings and related work can be divided into two categories: speaker-independent and speaker-dependent. The former represents most existing approaches, aiming to separate all speakers contained in the mixture. The latter ties the output to a specific speaker, which is more relevant to our framework. Finally, the effect of pitch priming is introduced, which motivates us to integrate pitch into the speech separation framework.

Speaker-Independent Speech Separation

Conventional deep-learning-based speaker separation models utilize multiple output layers in the network, each corresponding to one speaker. When it comes to speaker-independent models, the permutation problem needs to be addressed. Therefore, permutation invariant training (PIT) (Yu et al. 2017; Kolbaek et al. 2017) is proposed, which examines all possible label permutations for each utterance during training, and uses the one with the lowest utterance-level loss to train the separation network. However, PIT suffers from the output dimension mismatch problem since it assumes a fixed number of speakers. Deep clustering (DC) (Hershey et al. 2016) looks at the permutation problem from a different perspective. It firstly maps the time-frequency (T-F) units into an embedding space where a clustering algorithm (e.g., k-means) is performed to assign each T-F unit to one of the speakers in the mixture, which naturally tackles the output dimension mismatch problem meanwhile. However, the such clustering-based approach requires a certain number of clusters during evaluation, hence it cannot deal with the unknown speaker number in mixtures. Recently, Conv-TasNet (Luo and Mesgarani 2019) extends the first appearance of utterance-level PIT (uPIT) to the time domain using a convolutional encoder-decoder structure, which significantly improves the perceptual quality. Nowadays, dual-path network achieves a significant breakthrough. In (Yi Luo and Yoshioka 2020; Chen, Mao, and Liu 2020; Subakan et al. 2021), dual-path networks are chosen to apply intra- and inter-chunk operations iteratively.

Speaker-Dependent Speech Separation

For speaker-dependent separation, the information from the inferred speaker is taken as the condition to extract the corresponding speaker, i.e, separate one speaker each time, thus both permutation and speaker-number-unknown problems do not exist. VoiceFilter (Wang et al. 2019) separates the target speaker from multi-speaker signals, by using the speaker embedding of a reference signal from the target speaker. WaveSplit (Zeghidour and Grangier 2021) proposes an end-to-end source separation system by inferring a representation for each source and then estimating each source signal given the inferred representations. Shi et al. (Shi et al. 2020b) proposes a framework denoted as the Speaker-Conditional Chain Model, which first infers the identities of variable numbers of speakers from the mixture. Then, it takes the embedding from the inferred speakers sequentially as the condition to extract the corresponding speaker.

The proposed framework can be classified into this category, but we utilize pitch instead of speaker embedding as the prime, which corresponding mechanisms such as pitch priming can be found during auditory stream segregation in the multi-speaker scenario.

The Effect of Priming in ASA

There are two types of mechanisms for ASA (Bregman 1994): (1) primary mechanisms that process incoming mixtures of sounds in an automatic fashion using simple transformations, and (2) schema-based mechanisms that are more likely to be attention-, intention-, and knowledge-dependent. The former type is what the most existing frameworks are modeled on, while this work focused on the latter one. Recent research indicates that prior experience (e.g., priming) might be able to directly influence perception through non-attention-related

mechanisms (Snyder et al. 2012). In one early study (Dowling 1973), listeners were presented with two melodies at the same time. When the name of one of the tunes was given prior to hearing the interleaved melodies, it was easier to perceptually segregate it even when the two melodies were closer in pitch, demonstrating an effect of prior knowledge on perceptual segregation. In the sentence-level research, knowledge of the voice characteristics of the target talker could help the listener attend to the target speaker when other speakers are present, and improve the target speech recognition (Brungart 2001; Freyman, Balakrishnan, and Helfer 2004; Kidd Jr et al. 2005; Kidd Jr, Mason, and Gallun 2005). Experiments on voice-specific effects in auditory priming showed that the representation of F0 plays an important role, especially when speakers have similar pitch (Church and Schacter 1994). These findings inspire us to integrate the priming effect into nowadays speech separation framework, in order to address permutation and unknown speaker number problems.

PGSS Framework

Overview

The entire framework in Figure 1 is divided into two phases: pitch extraction and speech separation, according to two modeling processing in ASA. The pitch extraction phase models the process of bottom-up perceptual feature extraction. Given the input mixture, a two-stage approach is proposed to extract a list of pitch contours, each corresponding to one speaker. Any pitch contour of a specific speaker can be selected as the prime (e.g., speaker colored with yellow), then the speech separation phase aims to separate the corresponding speaker from the mixture using adversarial training, conditioned on the given pitch contour. The second phase models the effect of pitch priming in auditory stream segregation.

Phase I: Multi-Speaker Pitch Extraction

Pitch contour is a sequence of frequency values along the temporal dimension. It is often modeled through two processes: frame-level pitch candidates estimation and speaker assignment. The former aims to reliably estimate pitch candidates at each frame without indicating which speakers these pitch candidates belong to. Then the second process assigns the frame-level pitch candidates to specific speakers, to produce continuous pitch contours for each speaker. In practice, this process is associated with how to track the pitch candidates belonging to the same speaker across time. In this work, a novel two-stage multi-speaker pitch extraction approach is proposed, including *frame-level pitch estimation* and *utterance-level pitch tracking*.

Frame-Level Pitch Estimation The spectrum of a voiced sound is composed of a series of harmonics, regularly spaced in frequency at intervals of the fundamental frequency (F0), which is the principal determinant of perceived pitch (Middlebrooks et al. 2017). According to this characteristic, a method is proposed consisting of *harmonic modeling* and *harmonics-pitch mapping*.

Specifically, given a spectrogram X_m, which is transformed from the original mixture x in the time domain by

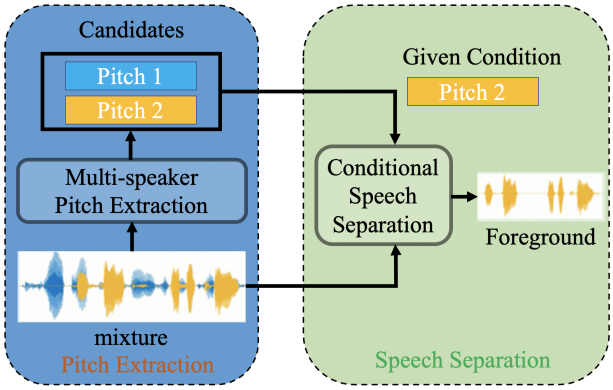

Figure 1: PGSS system overview.

Short-Time Fourier Transformation (STFT), where m is the frame index, this stage estimates a posterior pitch probability $P(z_m|X_m)$ where z_m denotes pitch states at frame m. We quantize the frequency range from 60 to 404 Hz into 67 bins using 24 bins per octave in a logarithmic scale (Liu and Wang 2018). Each bin corresponds to one pitch state. An additional pitch state represents silence or unvoiced speech, resulting in 68 pitch states (i.e., classes). $P(z_m(s)|X_m)$ equals one if the ground-truth pitch falls into the s-th bin. Therefore, frame-level pitch estimation can be treated as a multi-label classification task, which classifies the frame-level mixture input (no matter how many speakers it consists of) to a (several) specific pitch state(s) within 68 possible classes, hence this model is independent of the speaker number. Two-dimensional CNNs are applied first to capture harmonic structure and a fully-connected linear layer is followed to map the harmonic features to their corresponding pitches.

Cross-entropy is used as the loss function of this multi-label classification task, which is defined as:

$$L_m = \sum_{s=1}^{68} P(z_m(s)|X_m) log(O_m(s)) \quad (1)$$

where $O(\cdot)$ is the 68-units output layer followed by the sigmoid activation function.

Utterance-Level Pitch Tracking Given the frame-level result from last stage, this stage aims to link them belonging to the same speaker, to output multiple sequences of pitch contours, each for a specific speaker. It can be treated as a sequence-to-multi-sequence (seq2Mseq) mapping problem, where a conditional chain (cond-chain) model (Shi et al. 2020a) is adopted. Given the input sequence $O \in \mathcal{O}$, it is mapped to a set of N sequences $\mathbf{P} = \{\mathbf{p}_i | \in \{1, ..., N\}\}$, by a joint distribution of output sequences over an input sequence O as a product of conditional distributions. The model is supposed to automatically learn the efficient relationship between multiple output sequences, even though the relationship is mutually exclusive.

When implemented in our task, a conditional encoder-decoder structure is applied (Figure 2), where an additional $CondChain$ module preserves the information from previous output sequences and takes it as a condition for the

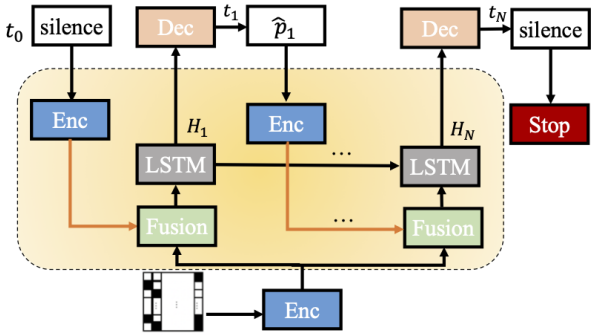

Figure 2: seq2Mseq mapping with conditional model for utterance-level pitch tracking. Blocks with same name share weights.

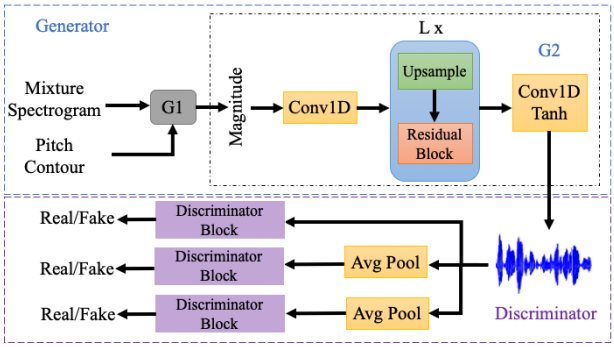

Figure 3: Pitch-CGAN architecture.

following outputs. In detail, input to the $CondChain$ is the *frame-level pitch estimation*. It can be seen as a one-hot matrix $O \in \mathbb{R}^{F,T}$, where F and T are frequency number and frame number, respectively, and the value 1/0 indicates whether there exists pitch or not within a T-F unit. The encoders (Enc) in $CondChain$ are shared across all steps. Each step corresponds to a process of decoding an output pitch sequence. $CondChain$ applies LSTM to store the information from previous sequences and regards them as conditions. For the $Fusion$ block, due to the same length of two inputs, a simple concatenation operation is used to stack them along the feature dimension for each frame. At each step i, the Decoder (Dec) is used to map the hidden state H_i into the final pitch contour $\hat{\mathbf{p}}_i \in \mathbb{R}^{68\times T}$. The whole process can be defined as:

$$\mathbf{E} = Enc(O) \in \mathbb{R}^{F^E \times T^E} \tag{2}$$

$$\mathbf{H}_i = CondChain(\mathbf{E}, \hat{\mathbf{p}}_{i-1}) \in \mathbb{R}^{F^C \times T^C} \tag{3}$$

$$\mathbf{D} = Dec(\mathbf{H}_i) \in \mathbb{R}^{68\times T} \tag{4}$$

Inspired by the conventional seq2seq model, we tackle the variable numbers of multiple sequences by predicting the end of the sequence symbol as a stop criterion. Specifically, an extra silent sequence with zero pitch values is attached at the end of output sequences, to satisfy the stop condition during training. The advantage of such stop criterion is that mixture data can consist of various numbers of speakers during training, and can be applied to the case of unknown numbers of speakers during inference.

Phase II: Pitch-Conditional Speech Separation

After obtaining a list of pitch contours from the last phase, then any of them can be selected as a condition to guide a specific speaker to be separated at this phase. To allow an explicit correspondence between pitch contour and mixture representation, we use a mixture spectrogram as the model input. Then in order to produce the target waveform directly, we are motivated by the MelGAN structure (Kumar et al. 2019), which is originally proposed for speech synthesis. The basic idea of MelGAN is it uses a stack of transposed convolutional layers to upsample the input spectrogram to its corresponding waveform. Phase information is supposed to be implicitly learned during this process.

Therefore, the final architecture is designed in a conditional generative adversarial network (CGAN) style (Mirza and Osindero 2014), shown in Figure 3. The selected pitch contour is treated as the condition, together with mixture spectrogram to form the input, denoted as Pitch-CGAN.

Generator The *Generator* (G) of Pitch-CGAN consists of two sub-modules: G1 and G2. G1 serves as a basic encoder to produce the target magnitude from the mixture, conditioned on a given pitch contour, while G2 applies MelGAN structure to synthesize the target waveform given the magnitude from G1.

In the G1, for a spectrogram $X \in \mathbb{R}^{T\times F}$ transformed from mixture x by STFT and a selected pitch contour $P \in \mathbb{R}^{T\times 68}$ where 68 is quantified frequency bins in section 3.2.1. To stack them along a new dimension, the frequency dimension should be the same. We follow the criterion of critical bands to extend the frequency dimension of P from 68 to F, and then stack it with X along a new channel dimension C to produce the input $Z \in \mathbb{R}^{2\times T\times F}$ to G1. G1 is composed of a U-Net-like CNN structure, followed by a bidirectional LSTM (BLSTM) and a final linear layer to map the output to the original frequency dimension. More details will be given in the supplementary.

In G2, a stack of transposed convolution is adopted to upsample the lower temporal resolution of the output magnitude from G1 to match that of the target waveform. Each transposed convolution is followed by a stack of residual blocks with dilated convolutions to increase the receptive field, as shown in the blue box in Figure 3.

Discriminator Same as the original MelGAN, multiple discriminators at different frequency scales are used, where each discriminator intends to learn features for different frequency ranges of speech. The multi-scale discriminators share the same network structure to operate on different speech scales in the frequency domain. Specifically, we adopt 3 discriminators ($D1$, $D2$, $D3$), where $D1$ operates on the scale of raw speech, whereas $D2$, $D3$ operate on raw speech downsam-

pled by a factor of 2 and 4 respectively. The downsampling is performed using strided average pooling with kernel size 4.

Objective Function With K discriminators, CGAN-Pitch conducts adversarial training with objectives as:

$$\min_{D_k} \mathbb{E}_y[(D_k(y|p)-1)^2] + \mathbb{E}_x[(D_k(G(X|p)))^2] \quad (5)$$

$$\min_{G} \mathbb{E}_x[\sum_{k=1}^{K}(D_k(G(X|p))-1)^2] \quad (6)$$

where D_k is the k-th discriminator, X represents the input mixture magnitude, p is the given pitch contour condition and y is the corresponding target waveform.

In addition, we adopt the multi-resolution STFT loss proposed in (Yang et al. 2021) to overcome the problem of measuring the differences between the potential features of true and fake speech and to make the convergence process faster. For a single STFT loss, we minimize the spectral convergence L_{sc} and log STFT magnitude L_{mag} between the target waveform y and the predicted audio $\tilde{y}$ from the generator $G(X|p)$:

$$L_{sc}(y,\tilde{y}) = \frac{\||STFT(y)| - |STFT(\tilde{y})\|_F}{\||STFT(y)|\|_F} \quad (7)$$

$$L_{mag\_G2}(y,\tilde{y}) = \frac{1}{N}\|log|STFT(y)| - log|STFT(\tilde{y})|\|_1 \quad (8)$$

where $\|\cdot\|_F$ and $\|\cdot\|_1$ are Frobenius and L1 norms, respectively, and N is the number of elements in the STFT magnitude. To restrict the output magnitude from G1, we minimize the log magnitude between the target magnitude S from G1 and STFT of the predicted audio $\tilde{y}$:

$$L_{mag\_G1}(G) = \mathbb{E}_{S,\tilde{y}}[\frac{1}{N}\|log|S| - log|STFT(\tilde{y})|\|_1] \quad (9)$$

The multi-resolution STFT objective function is an average of M single STFT losses with different analysis parameters (i.e., FFT size, window size and hop size), defined as:

$$L_{mr\_stft}(G) = \mathbb{E}_{y,\tilde{y}}[\frac{1}{M}\sum_{m=1}^{M}(L_{sc}^m(y,\tilde{y}) + L_{mag\_G2}^m(y,\tilde{y}))] \quad (10)$$

Hence the final training objectives of Pitch-CGAN is:

$$\min_{G} \mathbb{E}_x[\lambda \sum_{k=1}^{K}(D_k(G(x|p))-1)^2] + \mathbb{E}_x[L_{mr\_stft}(G)] + \mathbb{E}_x[L_{mag\_G1}(G)] \quad (11)$$

Experiments

Dataset and Setup

The proposed framework is evaluated on Wall Street Journal (WSJ0) corpus. The WSJ0-2mix and -3mix datasets are the benchmarks designed for speech separation, introduced by (Hershey et al. 2016). For WSJ0-2mix, the 30h training set and the 10h validation set contain two-speaker mixtures generated by randomly selecting speakers and utterances from the WSJ0 training set si_tr_s, and mixing them at various Signal-to-Noise Ratios (SNRs) uniformly chosen between 0 dB and 5 dB. The 5h test set was similarly generated using utterances from 18 speakers from the WSJ0 validation set si_dt_05 and evaluation set si_et_05. For three-speaker experiments, similar methods were adopted while the number of speakers was three, resulting in WSJ0-3mix.

Reference pitch is extracted from pre-mixed single-speaker utterances using the Praat (Boersma 2001). All data are sampled at 16 kHz. The input magnitudes are computed from STFT with 25 ms window length, 10 ms hop size, and the Hann window. Further details on the training setup are given in Appendix.

For the pitch extraction, the results are reported for both frame level (see Table 1) and utterance level via the error measure E_{Total} proposed in (Wohlmayr, Stark, and Pernkopf 2011). For speech separation, signal-to-distortion ratio (SDR), SDR improvement (SDRi) (Vincent, Virtanen, and Gannot 2018), perceptual evaluation of speech quality (PESQ) (Recommendation 2001) and short-time objective intelligibility (STOI) (Taal et al. 2011) are used in the ablation experiments (Gusó et al. 2022).

	Acc (%)	Prec (%)	Rec (%)
SG	98.5	91.6	84.7
DG	99.3	95.2	93.7
Overall	98.9	93.4	89.2

Table 1: Performance of frame-level pitch estimation w.r.t. different gender combinations.

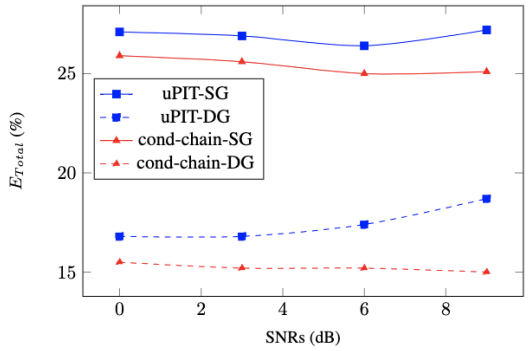

Figure 4: Performance of utterance-level pitch tracking w.r.t. different gender combinations, and absolute energy difference.

Pitch Extraction Performance

Frame-Level Pitch Estimation This stage actually performs a multi-label classification task, thus accuracy, precision and recall are used. Table 1 shows the results in terms of different gender combinations. Overall, the precision and recall are 93.4% and 89.2%, respectively, and performance for different-gender pairs (DG) is better than that for the same-

gender pairs (SG). It can be explained that, for same-gender speakers, the mean pitch (F0) of each speaker is relatively close, resulting in the overlap of harmonics that can not be resolved by the model.

To intuitively show the difference for different gender combinations, we plot the estimated frame-level pitch for SG and DG samples in the supplemental material. We compare them with corresponding reference pitch (sub-figure (a)) extracted from the pre-mixed single speaker through Praat. In general, the trend of estimated pitch variation is consistent with that of the reference, especially when the pitch contours are continuous over a long time span. However, for the segments where the duration of continuous pitch is relatively short, the model tends to connect them together or even miss them directly, such as the middle part for the SG pair. This situation is more terrible for SG mixtures, which is in our expectations as the harmonics for SG speakers are overlapped, and the model tends to produce only one pitch value.

Since speaker assignment has not been performed, we cannot plot the estimated results with different colors like the references.

Utterance-Level Pitch Tracking This stage tracks the frame-level results belonging to the same speaker, to produce a single utterance-level pitch contour for each speaker. The performance of the adopted cond-chain method is compared with a state-of-the-art (SOTA) baseline in speaker-independent pitch tracking, based on uPIT (uPIT-Pitch) (Liu and Wang 2018). The results are shown in Figure 4 for mixtures with different gender combinations and SNRs. It is obvious that for SD conditions, the error is systematically higher than that for DG conditions, due to the less accurate frame-level results from the last stage.

In general, the cond-chain method performs better than uPIT within each gender combination where cond-chain is more robust at different SNRs. Specifically, as the SNRs increase, E_{Total} first decreases and then increases. It seems to be inconsistent with the common sense that 0 dB is the most difficult situation. One possible explanation is that, when speakers are mixed at relatively high/low SNRs, someone's spectrum will be dominant while the others' will be masked. As uPIT produces pitch contours for all speakers simultaneously in a parallel style and only the information from input mixtures can be used, those masked speakers' might be less accurate, which degrades the overall performance. However, this situation is not observed in the cond-chain method. It may be caused that the cond-chain model predicts pitch contours successively where the previous pitch sequence can be used as a mutually exclusive condition for the current prediction. In addition, the proposed framework divides pitch extraction into two stages, each with a specific speech characteristics modeling, and optimized separately, resulting in better performance.

Speech Separation Performance

MelGAN module To verify the importance of MelGAN structure in the G2 of Pitch-CGAN, performance with and without G2 is shown in Table 2. The latter means the evaluated target waveform is reconstructed with the mixture phase,

Pitch-CGAN	SDRi (dB)	PESQ	STOI (%)
with G2	**15.3**	**3.4**	**94.3**
w/o G2	10.4	2.7	89.3

Table 2: Effect of MelGAN (G2) moduled.

Index	System	SDRi (dB)
F0	Pitch-CGAN	15.3
F1	+ Pre-tained G1	15.4
F2	+ MR STFT loss	15.6
F3	+ Deep ResStack	15.9

Table 3: Effect of different training strategies.

which significantly degrades the separation performance, especially on the perceptual metrics (i.e., STOI and PESQ). These two metrics are designed to reflect human perception of speech intelligibility and quality. While G2 is introduced, Pitch-CGAN can be optimized end to end to generate the target waveform directly, avoiding extra stage for phase estimation. Hence, Pitch-CGAN with and w/o G2 can be seen as the systems that represent frequency-domain and time-domain based CGAN models, respectively.

Training strategy Table 3 shows the SDRi results for Pitch-CGAN with different training strategies, each of them denoted as System F0-F3, respectively for simplicity. With G1 pre-trained, F1 outperforms the basic System F0 with a small increase in SDRi. Besides, the model converges much faster due to the restriction on the G1 output to be similar to the target magnitude first. When multi-resolution STFT loss is further introduced, performance is improved compared to F1. We assume that L_{mr_stft} can measure the difference between the true and fake speech in a lower signal level, while discriminator measures them on higher representations. Finally, by increasing the receptive field of F2 to become F3, we obtain a further improvement with the best SDRi among all the systems. With these applied tricks, we achieve better separation performance and training stability.

Compare with speaker-independent separation systems According to the output of generator, Pitch-CGAN w/o and with G2 module can be regarded as the systems in frequency domain and time domain, respectively. For Pitch-CGAN (w/o G2), it outperforms the corresponding uPIT (Kolbaek et al. 2017) and DPCL (Hershey et al. 2016) in all metrics. uPIT and DPCL are both mask-based systems, aiming at minimizing the mean square error (MSE) between the model output and reference. Such training objective may not be effective for all evaluation metrics. For example, MSE is more suitable for STOI and PESQ because they are calculated on magnitudes but not for SDR which is dependent on waveform signals. Pitch-CGAN applies the discriminator to judge the similarity between fake output and real reference, instead of specifically designed objectives for different evaluation metrics. However, since the final waveform is reconstructed by the mixture phase, the performance of frequency-domain

System	2-speaker			3-speaker		
	SDRi (dB)	PESQ	STOI (%)	SDRi (dB)	PESQ	STOI (%)
uPIT	9.4	2.6	87.7	4.7	2.1	79.2
DPCL	9.4	2.7	88.4	7.1	2.2	82.1
TasNet	11.1	2.9	90.7	9.6	2.4	84.6
Conv-TasNet	**15.6**	3.2	93.9	13.1	2.6	88.2
Pitch-CGAN (w/o G2)	10.4	2.7	89.3	8.9	2.2	84.9
Pitch-CGAN	15.3	**3.4**	**94.3**	**13.4**	**2.8**	**89.3**

Table 4: System comparison for speaker-independent separation.

System	SDRi (dB)
VoiceFilter (Wang et al. 2019)	7.0
TDAA (Shi et al. 2018)	7.5
SCCM (Shi et al. 2020b)	10.1
Pitch-CGAN	**15.3**

Table 5: System comparison for speaker-dependent separation.

systems is still far behind that of time-domain systems.

For Pitch-CGAN (with G2), generator outputs the waveform directly and skips the stage of target phase estimation, which achieves better performance than all the frequency-domain systems. For a 2-speaker mixture, Pitch-CGAN outperforms Conv-TasNet (Luo, Chen, and Mesgarani 2018) in STOI and PESQ but achieves a lower SDRi score. It can be attributed to the reason that ConvTasNet applies the objective function aiming at optimizing the SDRi directly, but the permutation problem still remains. To evaluate the generalization, results on WSJ0-3mix are also presented, where Pitch-CGAN achieves the highest scores in all metrics. Since only one output layer is set, corresponding to the conditional input, Pitch-CGAN can avoid the unknown speaker number problem.

Compare with speaker-dependent separation systems Different from the proposed system conditioned on pitch, existing speaker-conditional (dependent) separation systems rely on speaker embedding. VoiceFilter (Wang et al. 2019) uses an extra pre-trained speaker verification model to extract speaker embedding from reference signals, and then concatenates the embedding with a mixture spectrogram to produce the separated speech. Recently, Shi et al. proposes the Top-down auditory attention model (TDAA) (Shi et al. 2018) and Speaker-conditional chain model (SCCM) (Shi et al. 2020b) which both extract the embedding from mixture instead of reference signals.

Results of these systems are reported in Table 5. Pitch-CGAN outperforms all the speaker-embedding based systems. We summarize the following points. Firstly, pitch maintains the local patterns of spectrogram and variations along time dimension, while speaker embedding is the higher representation independent of time. For those speakers with similar timbre, pitch can remain discrimination. However, the error from pitch estimation may have more impact on separation performance than the latter, which can be regarded as the limitation. Secondly, as discussed before, existing speaker-dependent systems are all based on mask output, which degrades the performance as well.

Compare with SOTA separation systems Recently, monaural speech separation models have been extensively studied. Existing SOTA systems are mostly based on dual-path network, where different models are explored, from RNN-based DPRNN (Yi Luo and Yoshioka 2020) to currently transformer-based DPTNet (Chen, Mao, and Liu 2020) and SepFormer (Subakan et al. 2021). When comparing the proposed approach with these SOTA system, the following differences are listed: (1) As each time only one speaker is separated, according to the given pitch condition, the proposed method avoids the permutation and speaker number mismatch problems, which are still remained in SOTA systmes; (2) Our multi-stage approach has pros and cons (e.g., it is not end-to-end), but at least it improves the explainability of the system.

This work applies Conv-TasNet as the development architecture due to its efficiency and widely used in separation task. We are not aimed to replace the existing model as the SOTA one but to propose a conceptual and general separation framework which is motivated by the pitch-priming mechanism in ASA. Such framework would not be limited to a specific model architecture. We expect if the any SOTA architecture is adopted, the performance could be further improved. It is also a great opportunity to attract researchers' attention to the psychoacoustics community and to make more contribution to combine auditory mechanism with deep models.

Conclusion

In this work, we propose a novel pitch-guided speech separation (PGSS) framework, which is inspired by the effect of pitch priming in ASA mechanisms and addresses both permutation and unknown speaker number problems. Additionally, a separation approach based on CGAN is applied, leading to improved speech quality. However, such multi-stage approach has its cons which might be improved in the future by combining these two phases in an iterative style and updating each other. The SOTA model architectures should also be explored to evaluate the flexibility of this work.

Acknowledgments

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), and the High-performance Computing Platform of Peking University.

References

Bey, C.; and McAdams, S. 2002. Schema-based processing in auditory scene analysis. *Perception & psychophysics*, 64(5): 844–854.

Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9): 341–345.

Bregman, A. S. 1994. *Auditory scene analysis: The perceptual organization of sound*. MIT press.

Broadbent, D.; and Ladefoged, P. 1957. On the fusion of sounds reaching different sense organs. *The Journal of the Acoustical Society of America*, 29(6): 708–710.

Brown, G. J.; and Cooke, M. 1994. Computational auditory scene analysis. *Comput. Speech Lang.*, 8(4): 297–336.

Brungart, D. S. 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 109(3): 1101–1109.

Chen, J.; Mao, Q.; and Liu, D. 2020. Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation. In Meng, H.; Xu, B.; and Zheng, T. F., eds., *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, 2642–2646. ISCA.

Cherry, E. C. 1953. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5): 975–979.

Church, B. A.; and Schacter, D. L. 1994. Perceptual specificity of auditory priming: implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(3): 521.

Darwin, C.; and Hukin, R. 2000. Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *The Journal of the Acoustical Society of America*, 107(2): 970–977.

Darwin, C. J.; Brungart, D. S.; and Simpson, B. D. 2003. Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 114(5): 2913–2922.

Dowling, W. J. 1973. The perception of interleaved melodies. *Cognitive psychology*, 5(3): 322–337.

Freyman, R. L.; Balakrishnan, U.; and Helfer, K. S. 2004. Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *The Journal of the Acoustical Society of America*, 115(5): 2246–2256.

Glasberg, B. R.; and Moore, B. C. 1986. Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments. *The Journal of the Acoustical Society of America*, 79(4): 1020–1033.

Gusó, E.; Pons, J.; Pascual, S.; and Serrà, J. 2022. On loss functions and evaluation metrics for music source separation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 306–310. IEEE.

Hershey, J. R.; Chen, Z.; Roux, J. L.; and Watanabe, S. 2016. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, 31–35. IEEE.

Hu, G.; and Wang, D. L. 2010. A Tandem Algorithm for Pitch Estimation and Voiced Speech Segregation. *IEEE Trans. Speech Audio Process.*, 18(8): 2067–2079.

Hu, K.; and Wang, D. 2013. An Unsupervised Approach to Cochannel Speech Separation. *IEEE Trans. Speech Audio Process.*, 21(1): 120–129.

Kidd Jr, G.; Arbogast, T. L.; Mason, C. R.; and Gallun, F. J. 2005. The advantage of knowing where to listen. *The Journal of the Acoustical Society of America*, 118(6): 3804–3815.

Kidd Jr, G.; Mason, C. R.; and Gallun, F. J. 2005. Combining energetic and informational masking for speech identification. *The Journal of the Acoustical Society of America*, 118(2): 982–992.

Kolbaek, M.; Yu, D.; Tan, Z.; and Jensen, J. 2017. Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks. *IEEE ACM Trans. Audio Speech Lang. Process.*, 25(10): 1901–1913.

Kumar, K.; Kumar, R.; de Boissiere, T.; Gestin, L.; Teoh, W. Z.; Sotelo, J.; de Brébisson, A.; Bengio, Y.; and Courville, A. C. 2019. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 14881–14892.

Liu, Y.; and Wang, D. 2018. Permutation Invariant Training for Speaker-Independent Multi-Pitch Tracking. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, 5594–5598. IEEE.

Luo, Y.; Chen, Z.; and Mesgarani, N. 2018. Speaker-Independent Speech Separation With Deep Attractor Network. *IEEE ACM Trans. Audio Speech Lang. Process.*, 26(4): 787–796.

Luo, Y.; and Mesgarani, N. 2019. Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 27(8): 1256–1266.

McClelland, J. L.; Mirman, D.; and Holt, L. L. 2006. Are there interactive processes in speech perception? *Trends in cognitive sciences*, 10(8): 363–369.

Middlebrooks, J. C.; Simon, J. Z.; Popper, A. N.; and Fay, R. R. 2017. *The auditory system at the cocktail party*, volume 60. Springer.

Mirza, M.; and Osindero, S. 2014. Conditional Generative Adversarial Nets. *CoRR*, abs/1411.1784.

Recommendation, I.-T. 2001. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P. 862*.

Riecke, L.; Mendelsohn, D.; Schreiner, C.; and Formisano, E. 2009. The continuity illusion adapts to the auditory scene. *Hearing research*, 247(1): 71–77.

Riecke, L.; Micheyl, C.; Vanbussel, M.; Schreiner, C. S.; Mendelsohn, D.; and Formisano, E. 2011. Recalibration of the auditory continuity illusion: sensory and decisional effects. *Hearing research*, 277(1-2): 152–162.

Shamma, S. A.; Elhilali, M.; and Micheyl, C. 2011. Temporal coherence and attention in auditory scene analysis. *Trends in neurosciences*, 34(3): 114–123.

Shi, J.; Chang, X.; Guo, P.; Watanabe, S.; Fujita, Y.; Xu, J.; Xu, B.; and Xie, L. 2020a. Sequence to Multi-Sequence Learning via Conditional Chain Mapping for Mixture Signals. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Shi, J.; Xu, J.; Fujita, Y.; Watanabe, S.; and Xu, B. 2020b. Speaker-Conditional Chain Model for Speech Separation and Extraction. In Meng, H.; Xu, B.; and Zheng, T. F., eds., *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, 2707–2711. ISCA.

Shi, J.; Xu, J.; Liu, G.; and Xu, B. 2018. Listen, Think and Listen Again: Capturing Top-down Auditory Attention for Speaker-independent Speech Separation. In Lang, J., ed., *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, 4353–4360. ijcai.org.

Snyder, J. S.; Carter, O. L.; Hannon, E. E.; and Alain, C. 2009a. Adaptation reveals multiple levels of representation in auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, 35(4): 1232.

Snyder, J. S.; Carter, O. L.; Lee, S.-K.; Hannon, E. E.; and Alain, C. 2008. Effects of context on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, 34(4): 1007.

Snyder, J. S.; Gregg, M. K.; Weintraub, D. M.; and Alain, C. 2012. Attention, awareness, and the perception of auditory scenes. *Frontiers in psychology*, 3: 15.

Snyder, J. S.; Holder, W. T.; Weintraub, D. M.; Carter, O. L.; and Alain, C. 2009b. Effects of prior stimulus and prior perception on neural correlates of auditory stream segregation. *Psychophysiology*, 46(6): 1208–1215.

Snyder, J. S.; and Weintraub, D. M. 2011. Pattern specificity in the effect of prior Δf on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, 37(5): 1649.

Subakan, C.; Ravanelli, M.; Cornell, S.; Bronzi, M.; and Zhong, J. 2021. Attention Is All You Need In Speech Separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, 21–25. IEEE.

Taal, C. H.; Hendriks, R. C.; Heusdens, R.; and Jensen, J. 2011. An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech. *IEEE Trans. Speech Audio Process.*, 19(7): 2125–2136.

Takahashi, N.; Parthasaarathy, S.; Goswami, N.; and Mitsufuji, Y. 2019. Recursive speech separation for unknown number of speakers. *arXiv preprint arXiv:1904.03065*.

Vincent, E.; Virtanen, T.; and Gannot, S. 2018. *Audio source separation and speech enhancement*. John Wiley & Sons.

Wang, D.; and Brown, G. 2008. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. *IEEE Trans. Neural Networks*, 19(1): 199.

Wang, K.; Soong, F. K.; and Xie, L. 2019. A Pitch-aware Approach to Single-channel Speech Separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, 296–300. IEEE.

Wang, Q.; Muckenhirn, H.; Wilson, K. W.; Sridhar, P.; Wu, Z.; Hershey, J. R.; Saurous, R. A.; Weiss, R. J.; Jia, Y.; and Lopez-Moreno, I. 2019. VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking. In Kubin, G.; and Kacic, Z., eds., *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, 2728–2732. ISCA.

Wohlmayr, M.; Stark, M.; and Pernkopf, F. 2011. A Probabilistic Interaction Model for Multipitch Tracking With Factorial Hidden Markov Models. *IEEE Trans. Speech Audio Process.*, 19(4): 799–810.

Yang, G.; Yang, S.; Liu, K.; Fang, P.; Chen, W.; and Xie, L. 2021. Multi-Band Melgan: Faster Waveform Generation For High-Quality Text-To-Speech. In *IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, China, January 19-22, 2021*, 492–498. IEEE.

Yi Luo, Z. C.; and Yoshioka, T. 2020. Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, 46–50. IEEE.

Yu, D.; Kolbæk, M.; Tan, Z.-H.; and Jensen, J. 2017. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 241–245. IEEE.

Zeghidour, N.; and Grangier, D. 2021. Wavesplit: End-to-end speech separation by speaker clustering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 2840–2849.