

A UNIFIED ONE-SHOT PROSODY AND SPEAKER CONVERSION SYSTEM WITH SELF-SUPERVISED DISCRETE SPEECH UNITS

 preprint, 2022

Li-Wei Chen, Shinji Watanabe, Alexander Rudnicky

Carnegie Mellon University

INTRODUCTION

One-shot VC

Prosody Transfer

One-shot voice conversion is challenging as the model can only access source and target speech without speaker identities given.

Existing works mostly learned a speaker encoder jointly to isolate speaker information from prosody and language content.

However, in applications such as emotional style transfer, prosody needs to be controlled separately, which is ignored by most works.

The SpeechSplit and SRD-VC approaches gain the ability to individually control these attributes by decoupling speech into **prosody, speaker traits, and language content**.

Nevertheless, these methods do not take into account the correlation between prosody and content,

which leads to lower intelligibility of synthesized speech when the **prosody does not match the content**.

To solve this problem, this paper uses HuBERT's self-supervised labeling to assist the modeling of prosodic features,

successfully controlling various speech attributes independently while maintaining intelligibility.

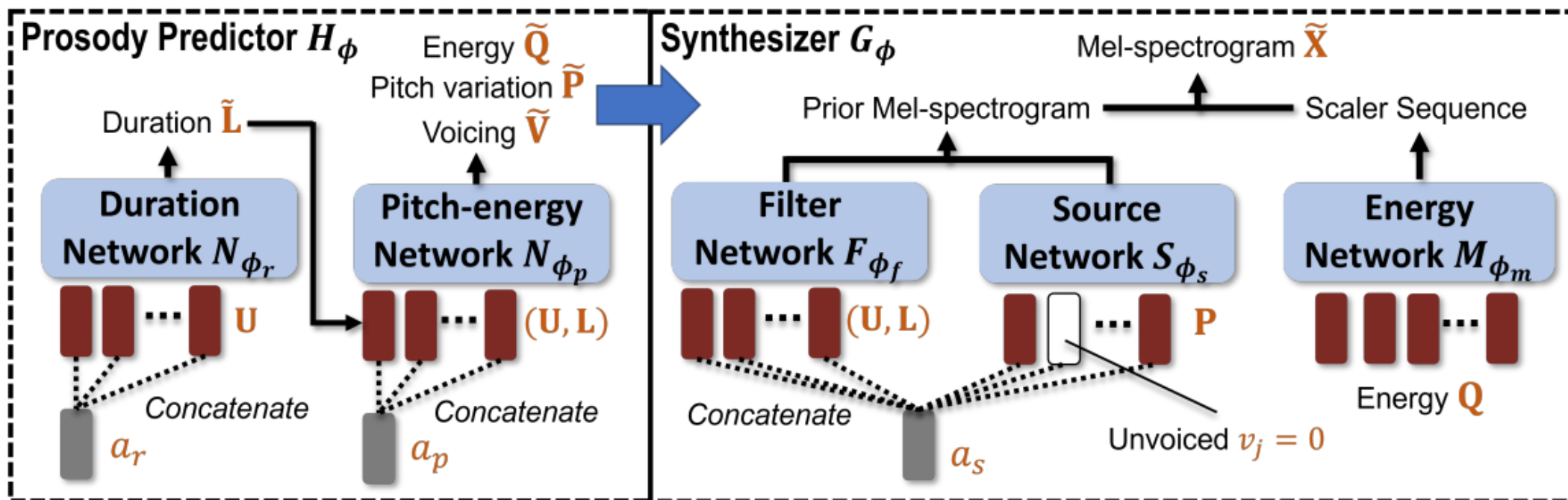
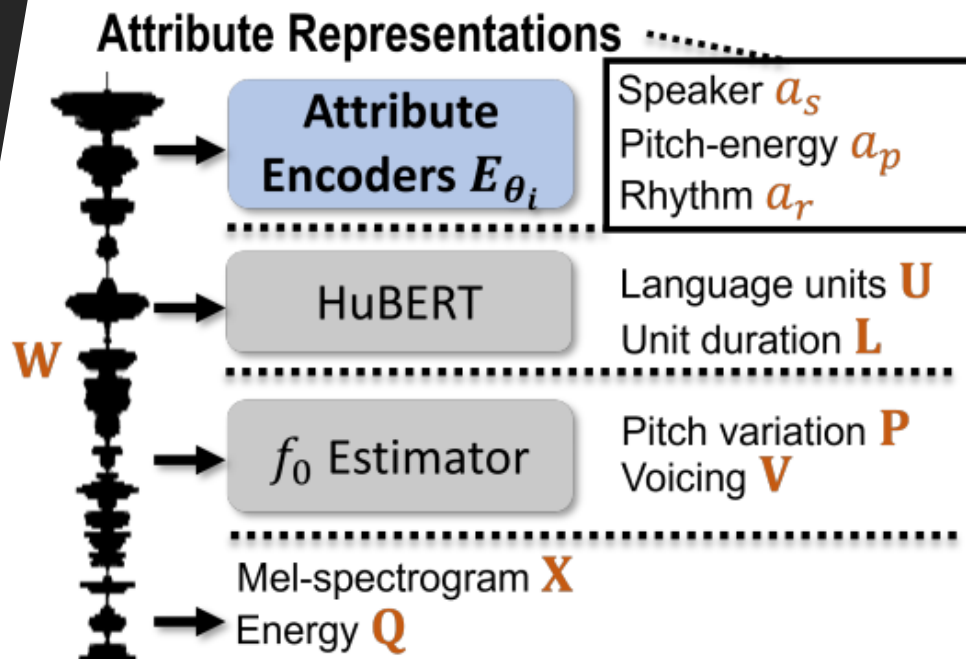
METHOD

$$a_i \leftarrow E_{\theta_i}(\mathbf{W})$$

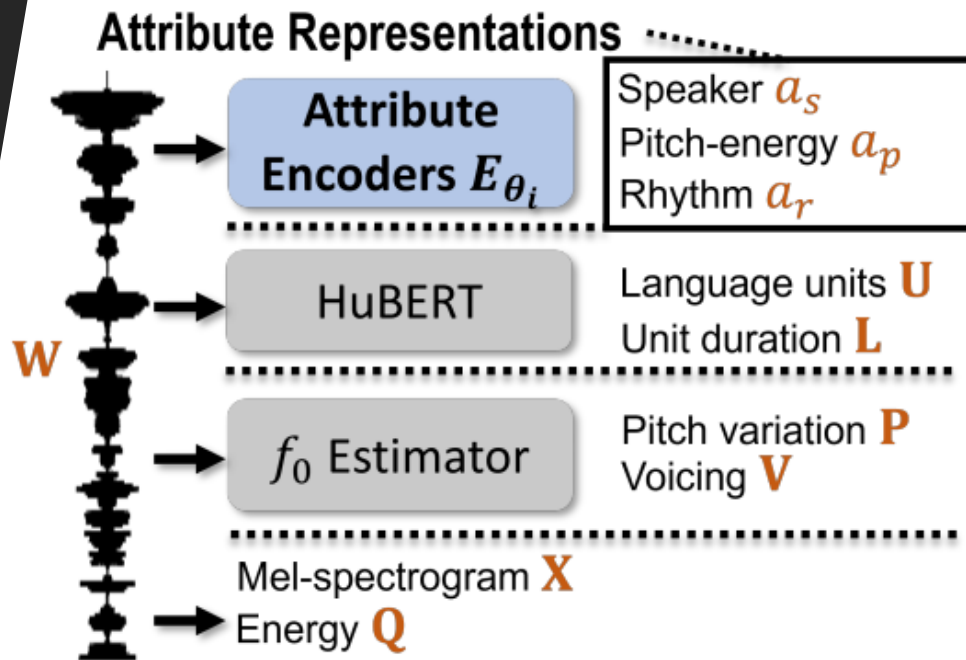
$$\{\tilde{\mathbf{P}}, \tilde{\mathbf{Q}}, \tilde{\mathbf{V}}, \tilde{\mathbf{L}}\} \leftarrow H_{\phi}(\mathbf{U}, a_p, a_r)$$

$$\tilde{\mathbf{X}} \leftarrow G_{\phi}(\mathbf{U}, \tilde{\mathbf{P}}, \tilde{\mathbf{Q}}, \tilde{\mathbf{V}}, \tilde{\mathbf{L}}, a_s)$$

System overview



Attribute Encoders

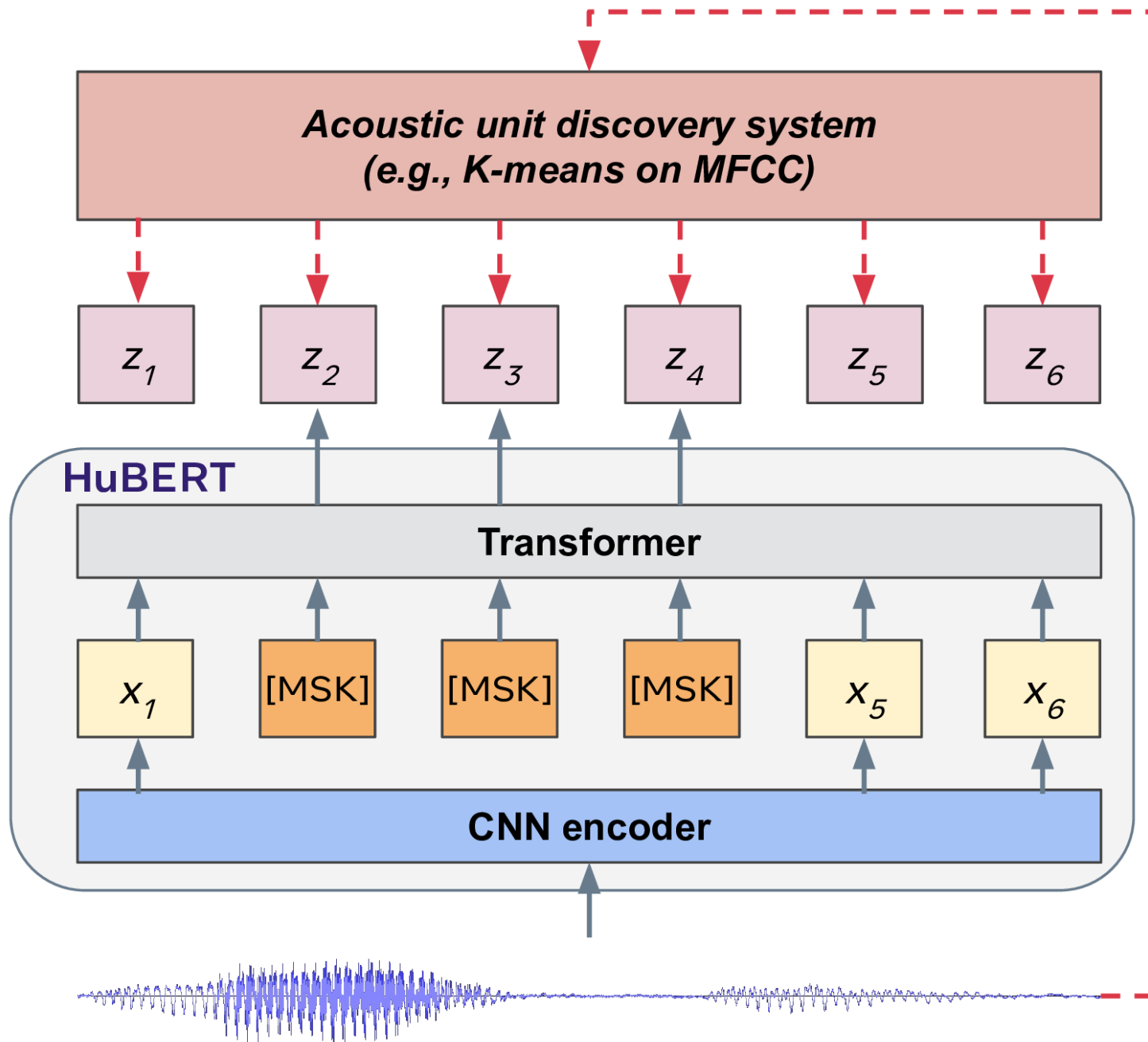


$\in \mathbb{R}^{d_a}$

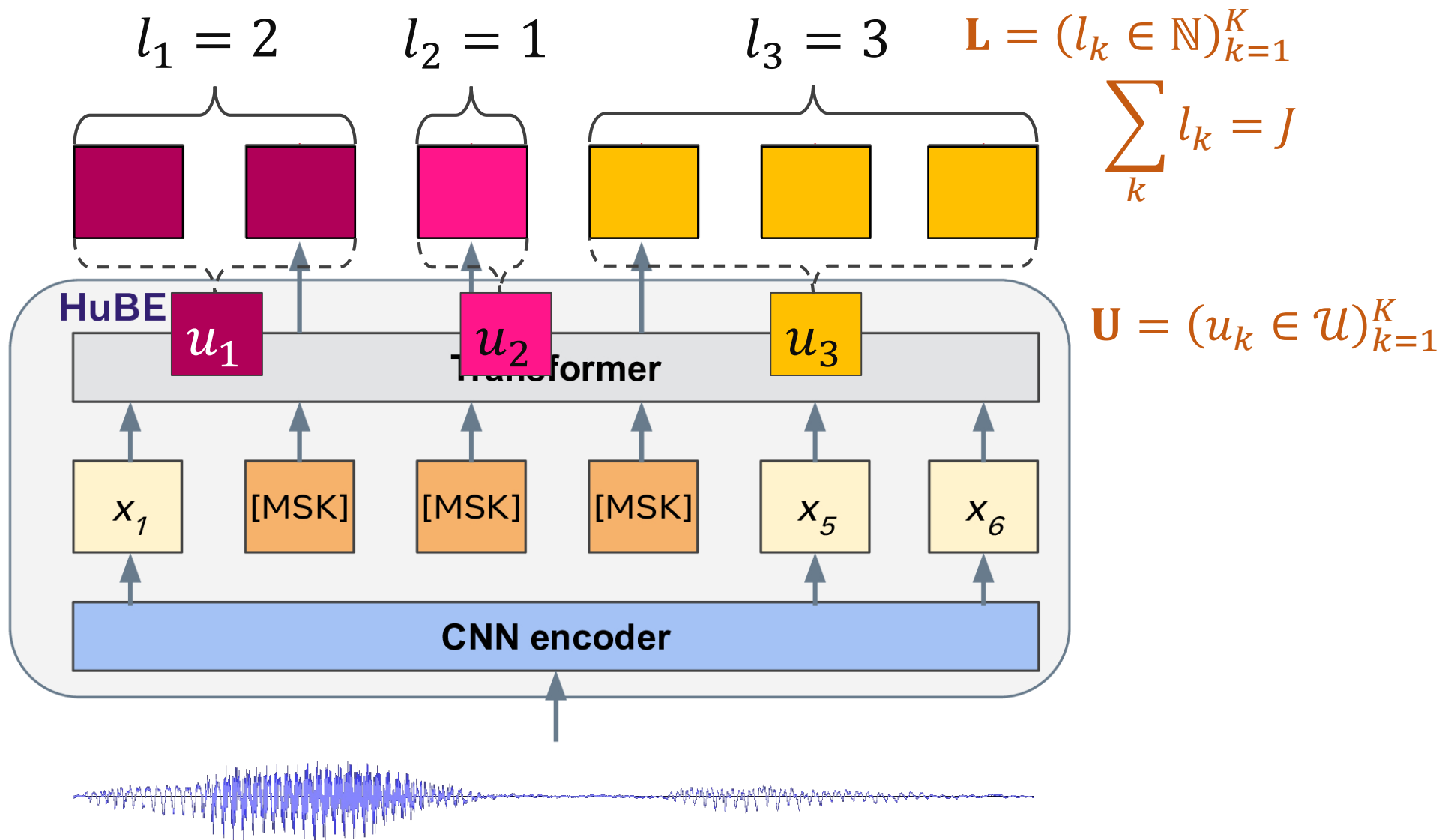
Adopt pretrained Wav2vec 2.0 as attribute encoders.

Fixed the CNN feature extractor, and fine-tune the 1-layer transformer separately for each encoder.

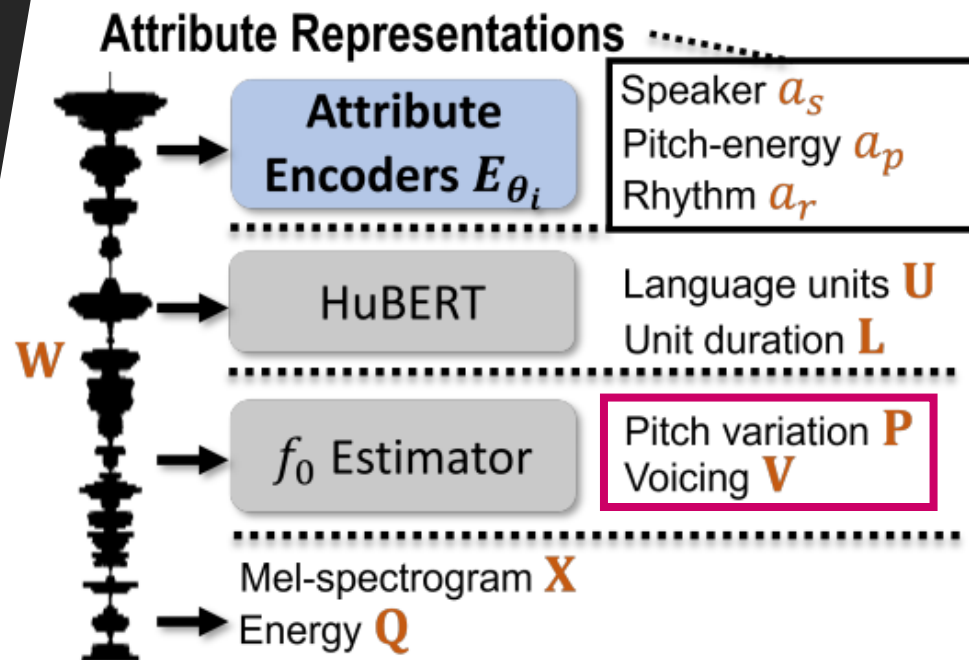
HuBERT



HuBERT



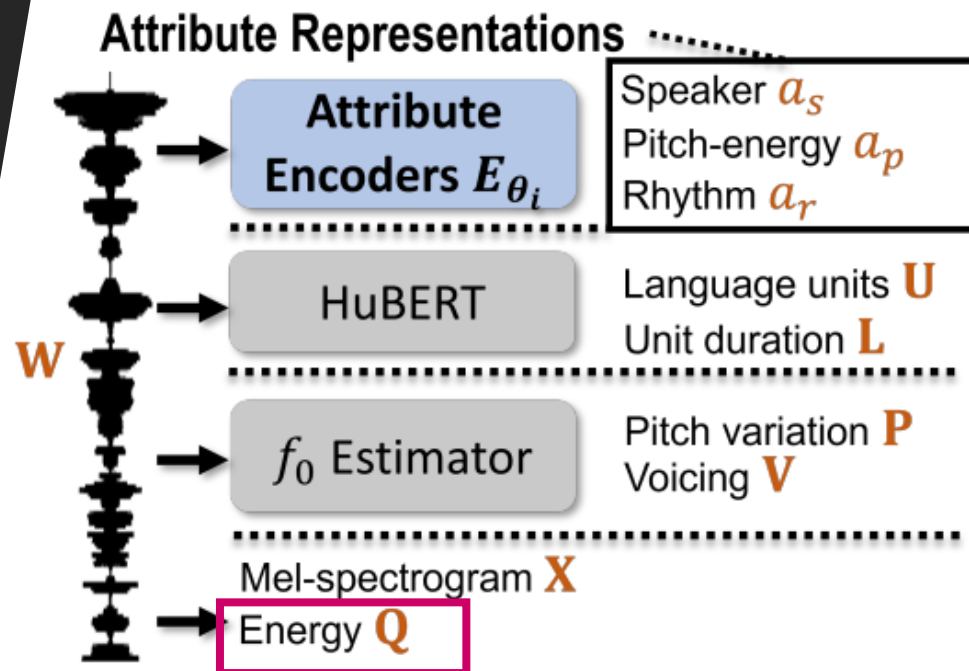
Pitch, Voicing & Energy



$\mathbf{P} = (p_j \in \mathbb{R})_{j=1}^J$ is mean-normalized pitch in Hertz.

$\mathbf{V} = (v_j \in \{0,1\})_{j=1}^J$, 0 represents unvoiced and 1 represents voiced.

Obtain \mathbf{P} and \mathbf{V} from pitch estimator such as CREPE, YAPPT.



$\mathbf{P} = (p_j \in \mathbb{R})_{j=1}^J$ is mean-normalized pitch in Hertz.

$\mathbf{V} = (v_j \in \{0,1\})_{j=1}^J$, 0 represents unvoiced and 1 represents voiced.

Obtain \mathbf{P} and \mathbf{V} from pitch estimator such as CREPE, YAPPT.

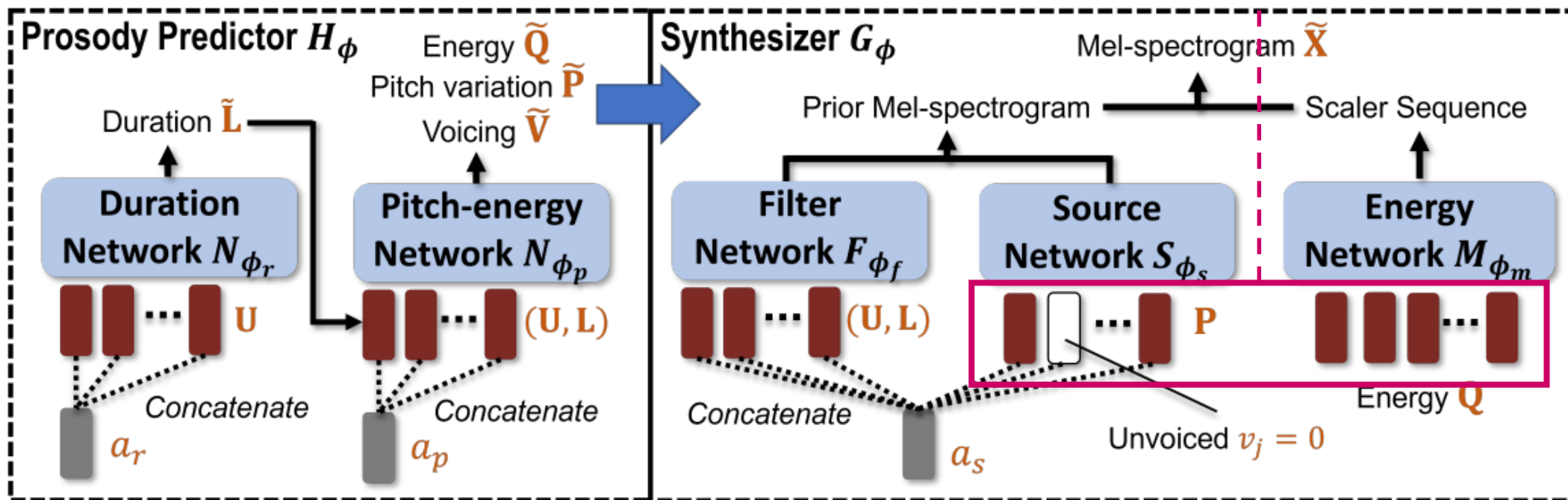
Obtain energy $\mathbf{Q} = (q_n \in \mathbb{R})_{n=1}^N$ from the frame-wise L2 norm of the linear spectrogram.

Convert **P, Q** to representation.

1. Scalar to bin weight

$$b(h_j) = \{b_i(h_j)\}_{i=1}^B, b_i(h_j) = \exp\left(-\frac{(h_j - c_i)^2}{2\sigma^2}\right)$$

$$c_i = i\ell_b + p_{min}, h \in \{p, q\}$$

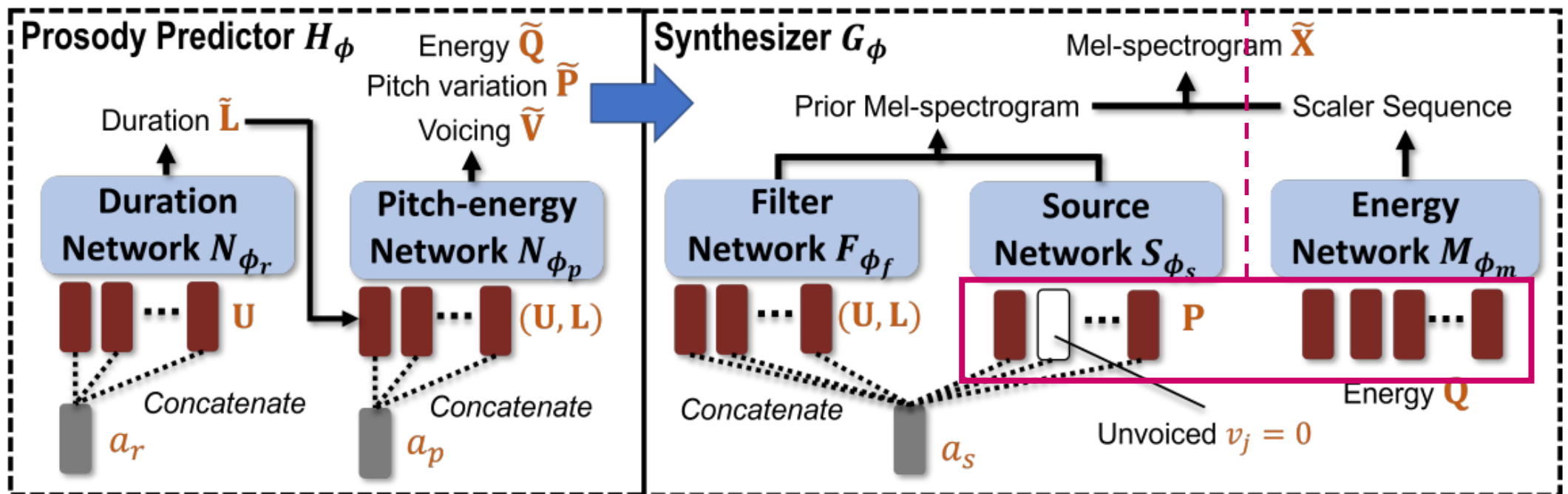


Convert **P, Q** to representation.

2. Weighted sum with learnable embedding

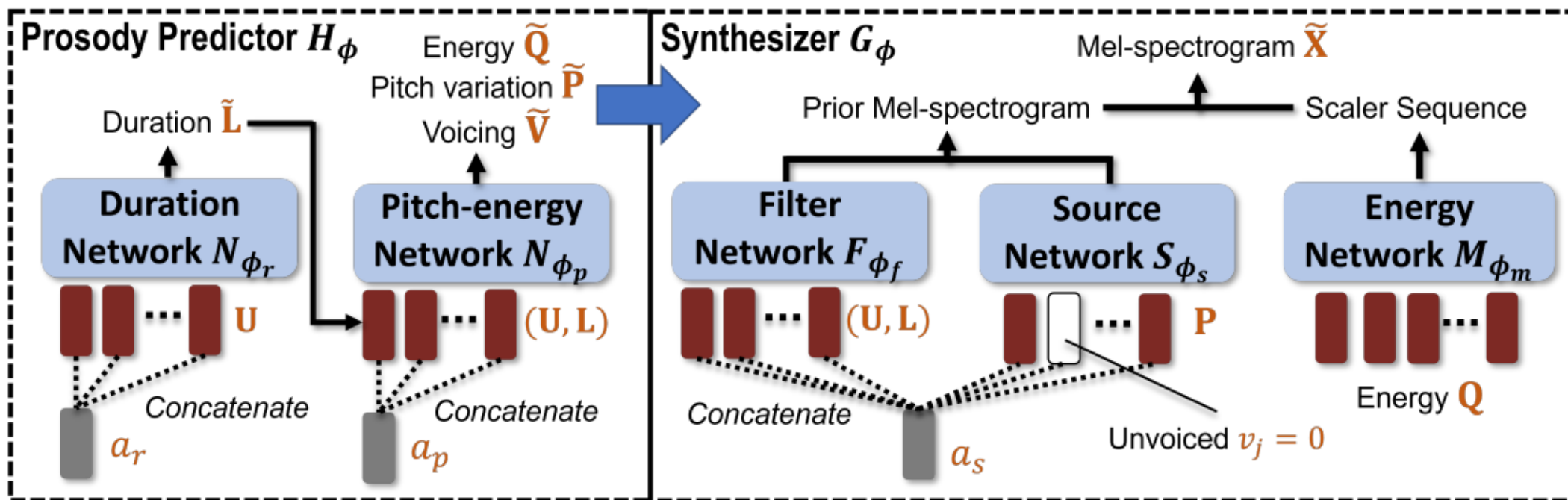
$$o(h_j, E^h) = \frac{\sum_{i=1}^B b_i(p_j) e_i^h}{\sum_{i=1}^B b_i(p_j)}, E^h = \{e_i^h \in \mathbb{R}^{d_e}\}_{i=0}^B$$

3. replace the unvoiced frames (j where $v_j = 0$) of with $o(p_j, E^p)$ another learnable embedding.



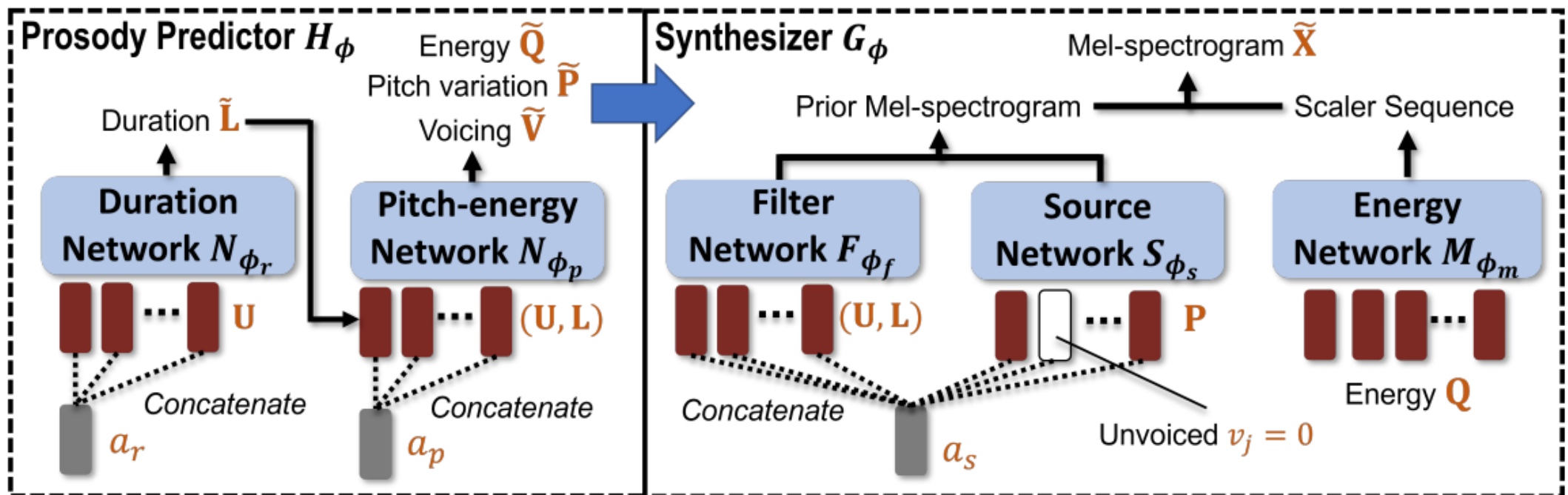
$$\tilde{\mathbf{X}} \leftarrow G_{\phi}(\mathbf{U}, \tilde{\mathbf{P}}, \tilde{\mathbf{Q}}, \mathbf{V}, \mathbf{L}, \mathbf{a}_s)$$

the poor estimation in the early stage of training is not conducive to G_{ϕ} considering prosody information, which will cause G_{ϕ} to ignore prosody information.



$$\tilde{\mathbf{X}} \leftarrow G_{\phi}(\mathbf{U}, \tilde{\mathbf{P}}, \tilde{\mathbf{Q}}, \mathbf{V}, \mathbf{L}, a_s)$$

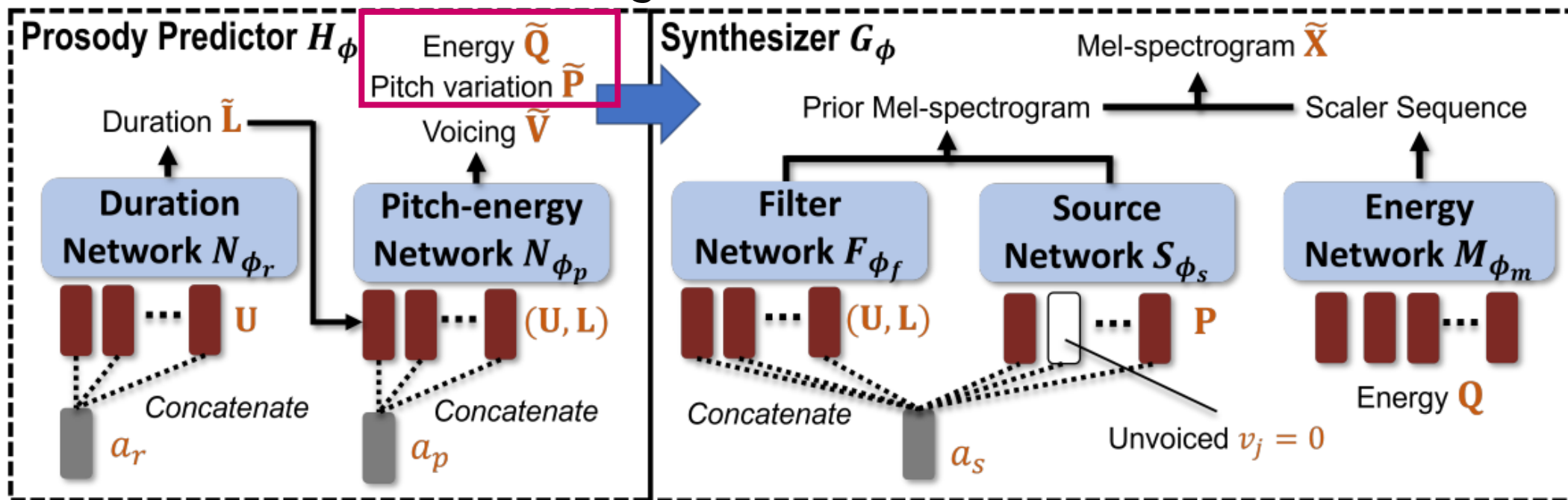
$$\tilde{\mathbf{X}} \leftarrow G_{\phi}\left(\mathbf{U}, \frac{b(\mathbf{P}) + \tilde{\mathbf{P}}}{2}, \frac{b(\mathbf{Q}) + \tilde{\mathbf{Q}}}{2}, \mathbf{V}, \mathbf{L}, a_s\right)$$



$$\tilde{\mathbf{X}} \leftarrow G_{\phi} \left(\mathbf{U}, \frac{b(\mathbf{P}) + \tilde{\mathbf{P}}}{2}, \frac{b(\mathbf{Q}) + \tilde{\mathbf{Q}}}{2}, \mathbf{V}, \mathbf{L}, a_s \right)$$

$$\mathcal{L}_{mel}(\mathbf{X}, \tilde{\mathbf{X}}) + \mathcal{L}_E(\mathbf{Q}, \tilde{\mathbf{Q}}) + \mathcal{L}_P(\mathbf{P}, \tilde{\mathbf{P}}) + \mathcal{L}_V(\mathbf{V}, \tilde{\mathbf{V}}) + \mathcal{L}_L(\mathbf{L}, \tilde{\mathbf{L}}) + \mathcal{L}_{GAN}(\tilde{\mathbf{X}})$$

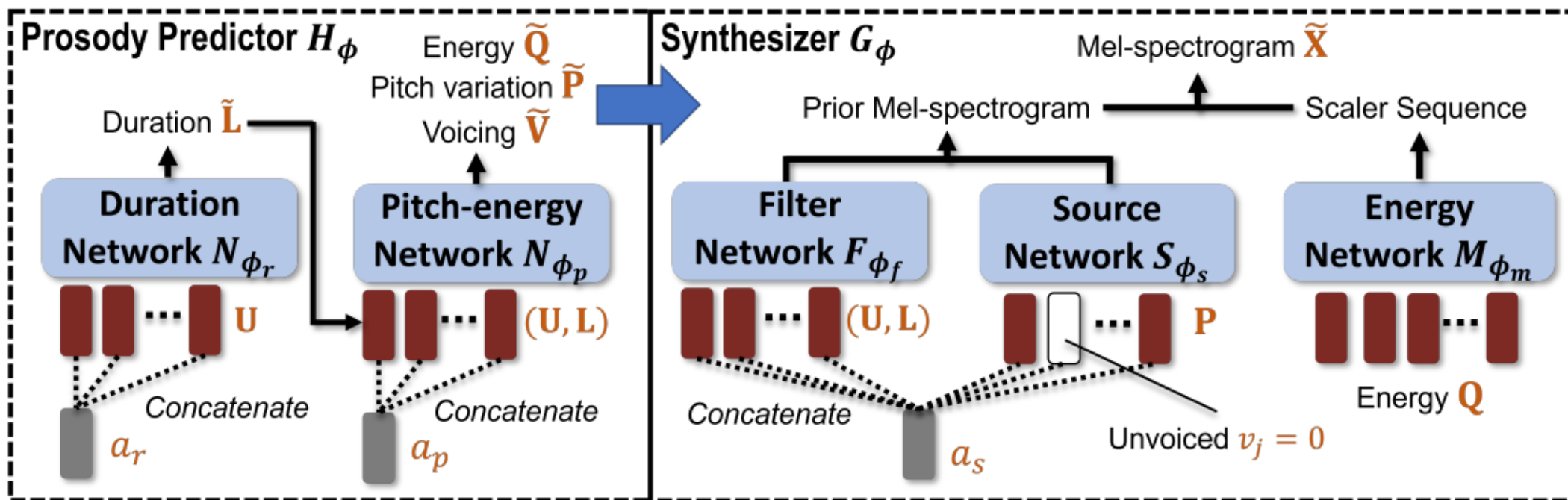
Predict bin weight



$$\tilde{\mathbf{X}} \leftarrow G_{\phi} \left(\mathbf{U}, \frac{b(\mathbf{P}) + \tilde{\mathbf{P}}}{2}, \frac{b(\mathbf{Q}) + \tilde{\mathbf{Q}}}{2}, \mathbf{V}, \mathbf{L}, a_s \right)$$

$$\mathcal{L}_{mel}(\mathbf{X}, \tilde{\mathbf{X}}) + \mathcal{L}_E(\mathbf{Q}, \tilde{\mathbf{Q}}) + \mathcal{L}_P(\mathbf{P}, \tilde{\mathbf{P}}) + \mathcal{L}_V(\mathbf{V}, \tilde{\mathbf{V}}) + \mathcal{L}_L(\mathbf{L}, \tilde{\mathbf{L}}) + \mathcal{L}_{GAN}(\tilde{\mathbf{X}})$$

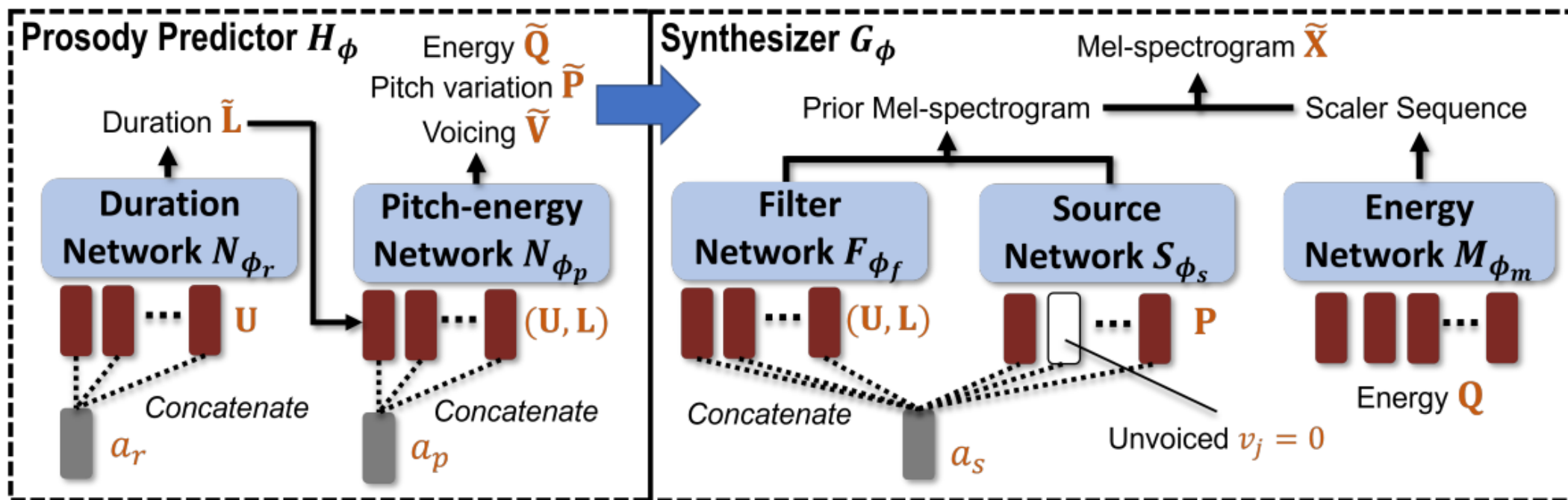
L1 loss



$$\tilde{\mathbf{X}} \leftarrow G_{\phi} \left(\mathbf{U}, \frac{b(\mathbf{P}) + \tilde{\mathbf{P}}}{2}, \frac{b(\mathbf{Q}) + \tilde{\mathbf{Q}}}{2}, \mathbf{V}, \mathbf{L}, a_s \right)$$

$$\mathcal{L}_{mel}(\mathbf{X}, \tilde{\mathbf{X}}) + \mathcal{L}_E(\mathbf{Q}, \tilde{\mathbf{Q}}) + \mathcal{L}_P(\mathbf{P}, \tilde{\mathbf{P}}) + \mathcal{L}_V(\mathbf{V}, \tilde{\mathbf{V}}) + \mathcal{L}_L(\mathbf{L}, \tilde{\mathbf{L}}) + \mathcal{L}_{GAN}(\tilde{\mathbf{X}})$$

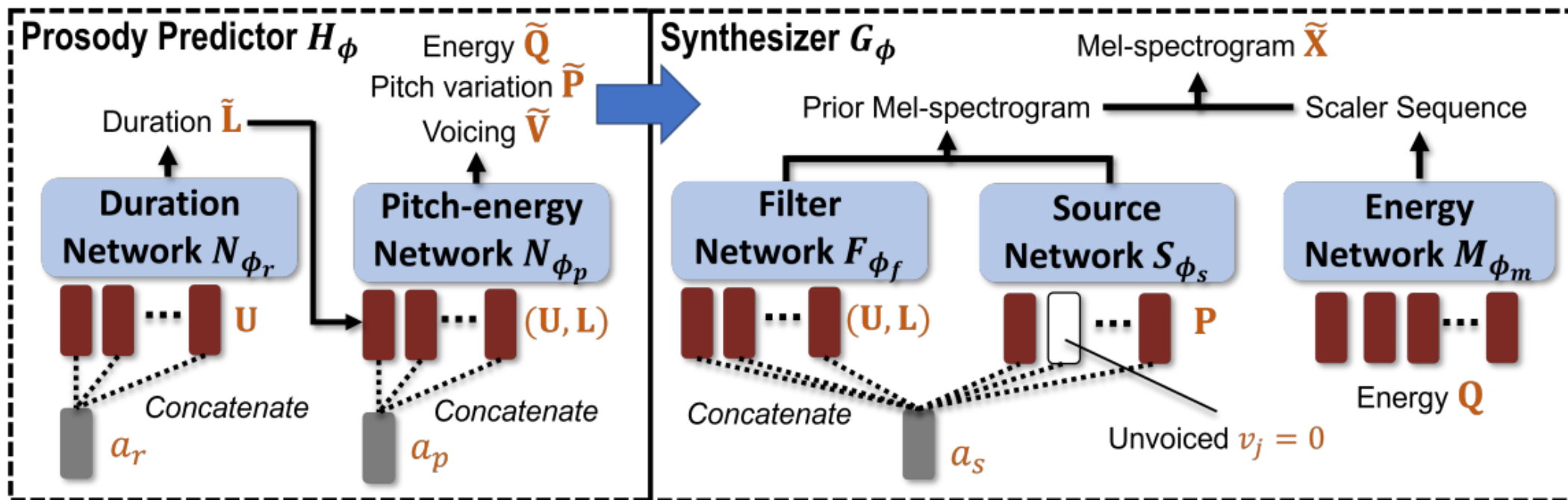
BCE loss of bin &
MSE loss of representation



$$\tilde{\mathbf{X}} \leftarrow G_{\phi} \left(\mathbf{U}, \frac{b(\mathbf{P}) + \tilde{\mathbf{P}}}{2}, \frac{b(\mathbf{Q}) + \tilde{\mathbf{Q}}}{2}, \mathbf{V}, \mathbf{L}, a_s \right)$$

$$\mathcal{L}_{mel}(\mathbf{X}, \tilde{\mathbf{X}}) + \mathcal{L}_E(\mathbf{Q}, \tilde{\mathbf{Q}}) + \mathcal{L}_P(\mathbf{P}, \tilde{\mathbf{P}}) + \mathcal{L}_V(\mathbf{V}, \tilde{\mathbf{V}}) + \mathcal{L}_L(\mathbf{L}, \tilde{\mathbf{L}}) + \mathcal{L}_{GAN}(\tilde{\mathbf{X}})$$

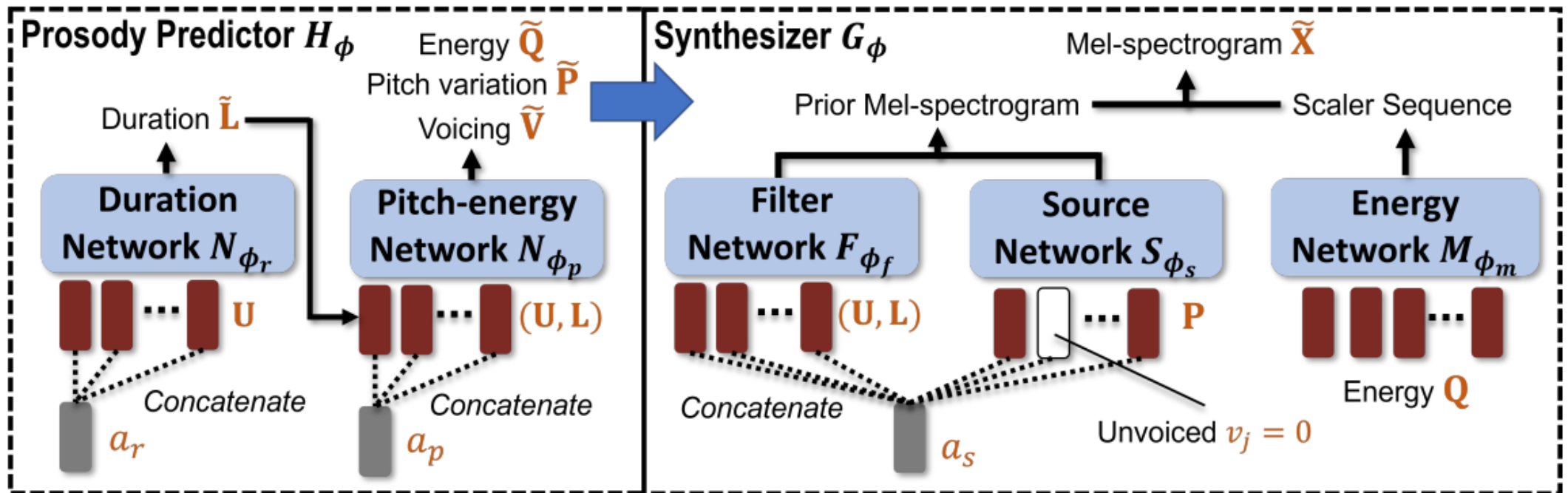
BCE loss



$$\tilde{\mathbf{X}} \leftarrow G_{\phi} \left(\mathbf{U}, \frac{b(\mathbf{P}) + \tilde{\mathbf{P}}}{2}, \frac{b(\mathbf{Q}) + \tilde{\mathbf{Q}}}{2}, \mathbf{V}, \mathbf{L}, a_s \right)$$

$$\mathcal{L}_{mel}(\mathbf{X}, \tilde{\mathbf{X}}) + \mathcal{L}_E(\mathbf{Q}, \tilde{\mathbf{Q}}) + \mathcal{L}_P(\mathbf{P}, \tilde{\mathbf{P}}) + \mathcal{L}_V(\mathbf{V}, \tilde{\mathbf{V}}) + \boxed{\mathcal{L}_L(\mathbf{L}, \tilde{\mathbf{L}})} + \mathcal{L}_{GAN}(\tilde{\mathbf{X}})$$

MSE loss



RESULTS

Train on the VCTK corpus,
and evaluate on the LibriTTS corpus.

Table 1: Evaluation results for **unseen speaker transfer**. The right columns are naturalness and speaker similarity MOS with 95% CI. *GT* stands for the ground truth speech.

Method	CER ↓	SES ↑	Naturalness ↑	Similarity ↑
<i>GT</i>	5.5%	<i>n/a</i>	3.95 ± 0.11	<i>n/a</i>
<i>AutoVC</i>	88.4%	0.14	2.58 ± 0.17	2.62 ± 0.15
<i>SRDVC</i>	34.7%	0.17	3.25 ± 0.15	2.56 ± 0.14
<i>Ours</i>	7.5%	0.34	3.50 ± 0.14	2.80 ± 0.14
(-W2V2)	7.8%	0.27	3.47 ± 0.13	2.62 ± 0.15
(-Joint Opt.)	7.3%	0.32	3.55 ± 0.14	2.68 ± 0.14

SES: Speaker Embedding Similarity

Table 2: Average Pearson correlation coefficients (PCC) of pitch and energy between target and speaker-converted speech.

PCC	<i>AutoVC</i>	<i>SRDVC</i>	<i>Ours</i> ($\tilde{\mathbf{P}}, \tilde{\mathbf{Q}}$)	<i>Ours</i> (\mathbf{P}, \mathbf{Q})
$\log f_0$	0.09	0.49	0.30	0.51
Energy	0.05	0.82	0.78	0.91

Table 3: Evaluation results for **prosody transfer**. The right columns are naturalness and prosody similarity MOS with 95% CI.

Method	CER ↓	EES ↑	Naturalness ↑	Similarity ↑
<i>GT</i>	5.5%	<i>n/a</i>	3.95 ± 0.11	<i>n/a</i>
<i>SRDVC</i>	49.9%	0.28	3.06 ± 0.16	2.58 ± 0.15
<i>Ours</i>	8.9%	0.42	3.49 ± 0.13	2.69 ± 0.14
(-W2V2)	10.1%	0.35	3.39 ± 0.14	2.57 ± 0.15
(-Joint Opt.)	9.0%	0.38	3.36 ± 0.14	2.50 ± 0.14

EES: Emotion Embedding Similarity

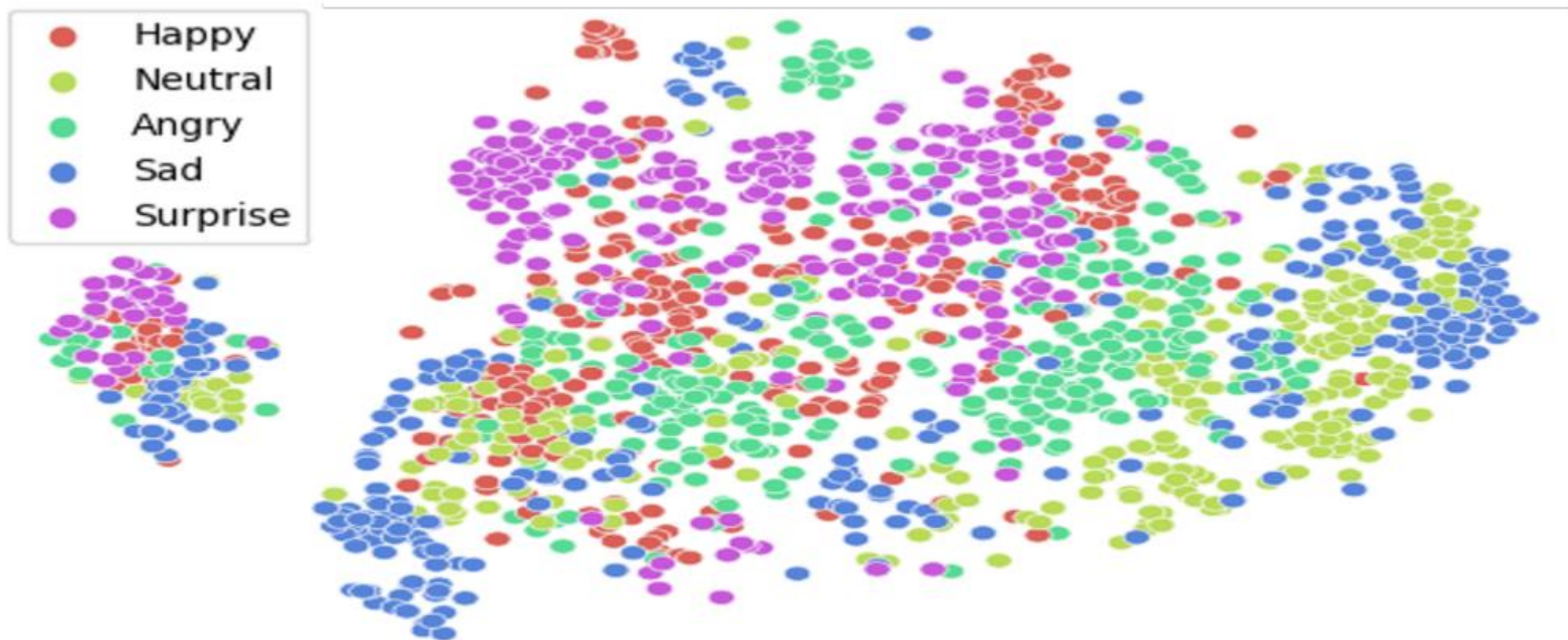


Fig. 2: Visualization of a_p (pitch-energy) by running t-SNE on ESD.

CONCLUSION

A new VC framework is proposed to solve the problem of speech decomposition.

Language content can be isolated without transcribed text using HuBERT.

When content is isolated, rhythm information can be decomposed from speech.

A new VC framework
solve the problem
decomposition.

When cor
be decom



It can be isolated
text using HuBERT.

Information can