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

arXiv preprint, 2022

Li-Wei Chen, Shinji Watanabe, Alexander Rudnicky

Carnegie Mellon University

INTRODUCTION

One-shot VC

Prosody Transfer

INTRODUCTION

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.

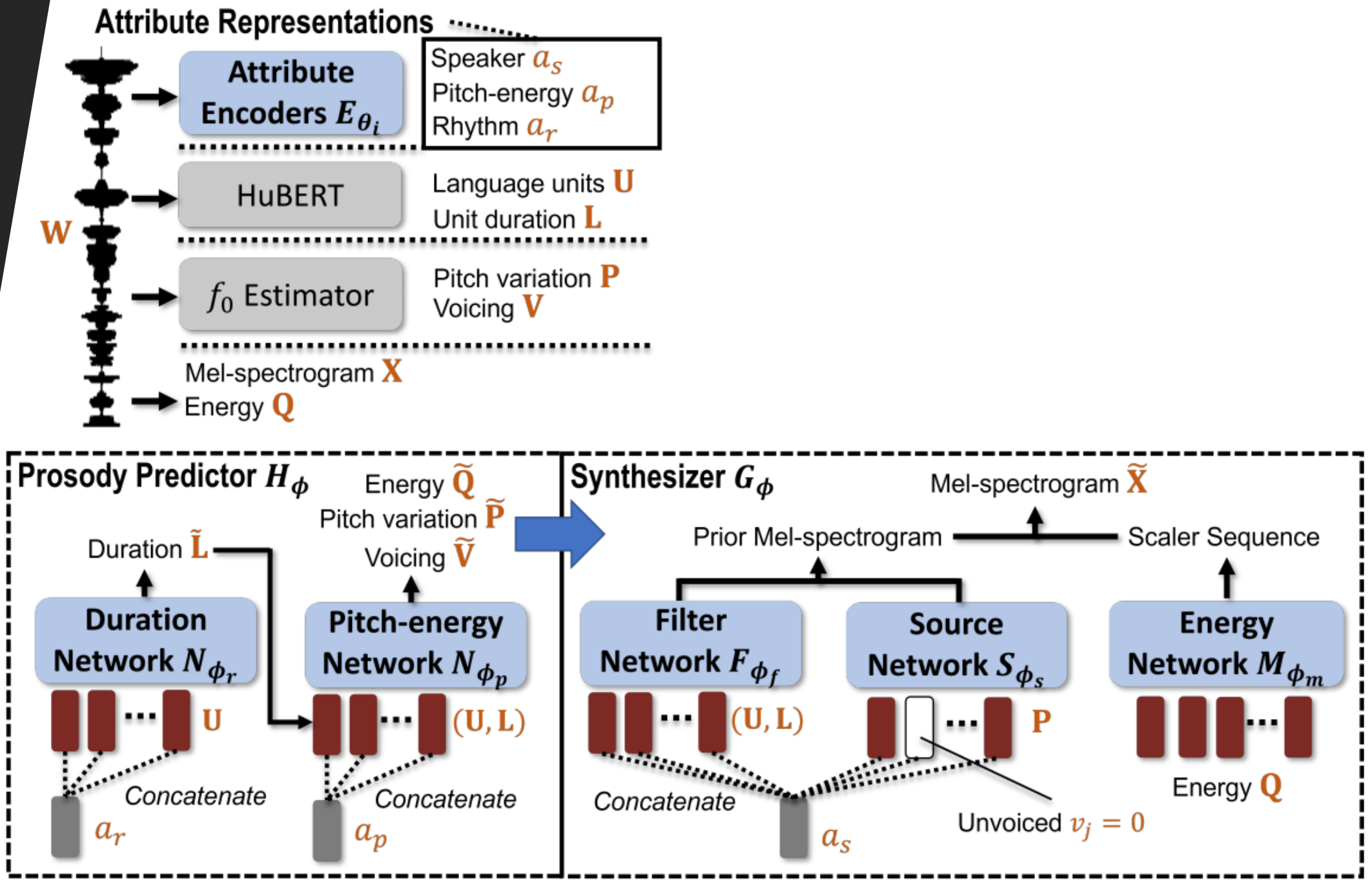
METHOD

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

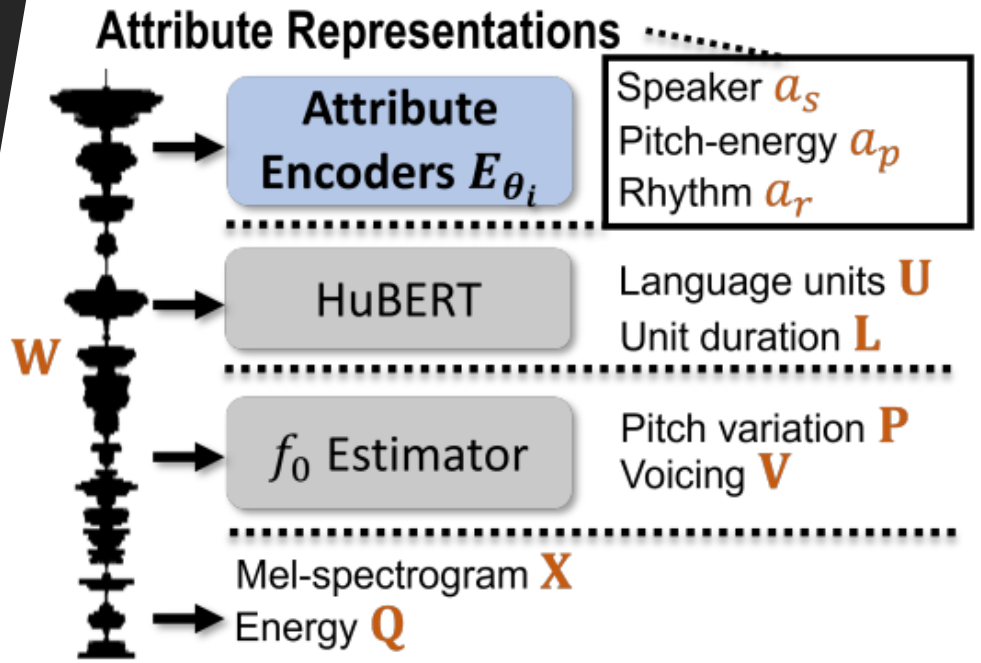
$$\{\widetilde{\mathbf{P}}, \widetilde{\mathbf{Q}}, \widetilde{\mathbf{V}}, \tilde{\mathbf{L}}\} \leftarrow H_{\phi}(\mathbf{U}, a_p, a_r)$$

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

System overview



Attribute Encoders

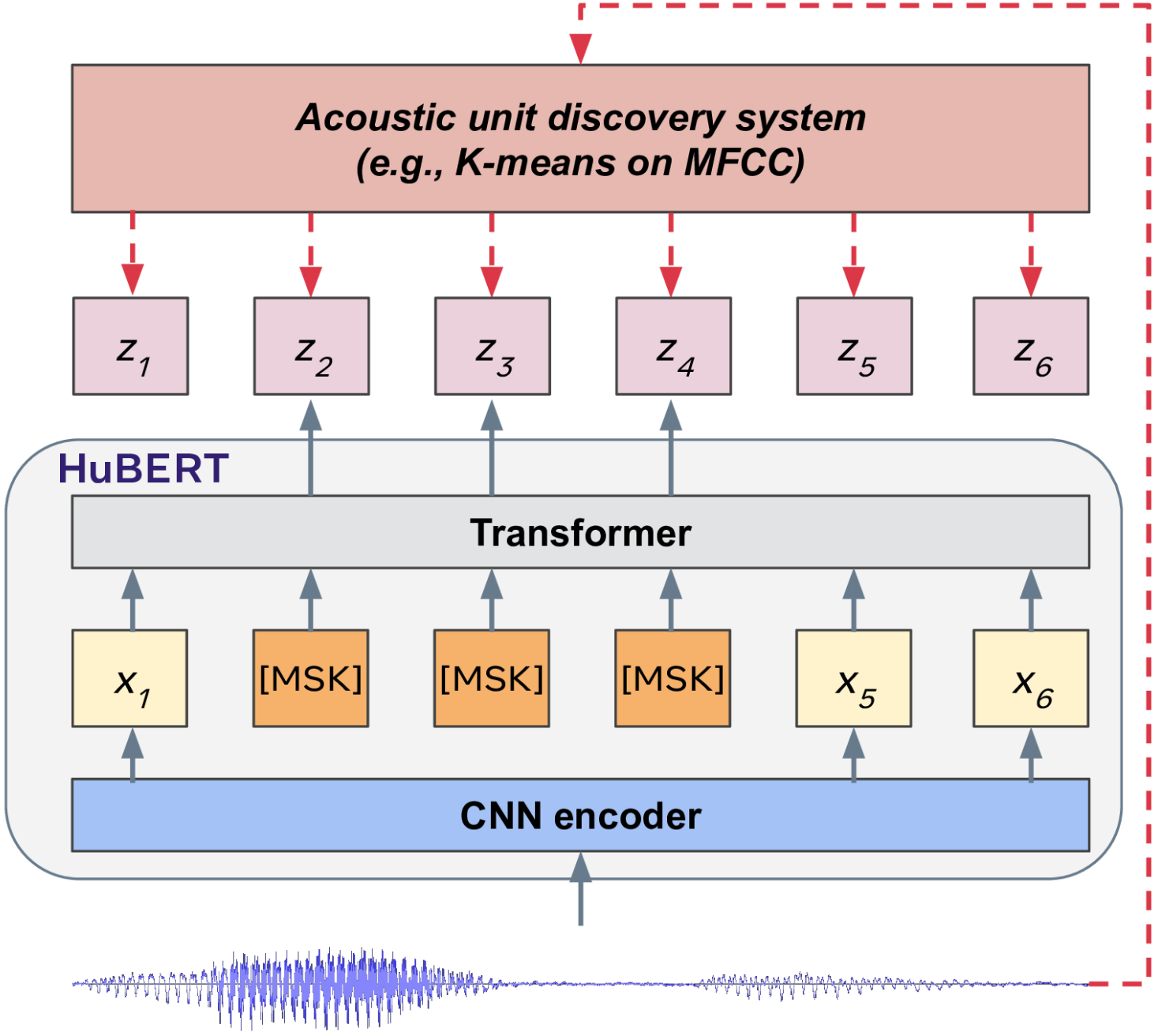


$\in \mathbb{R}^{d_a}$

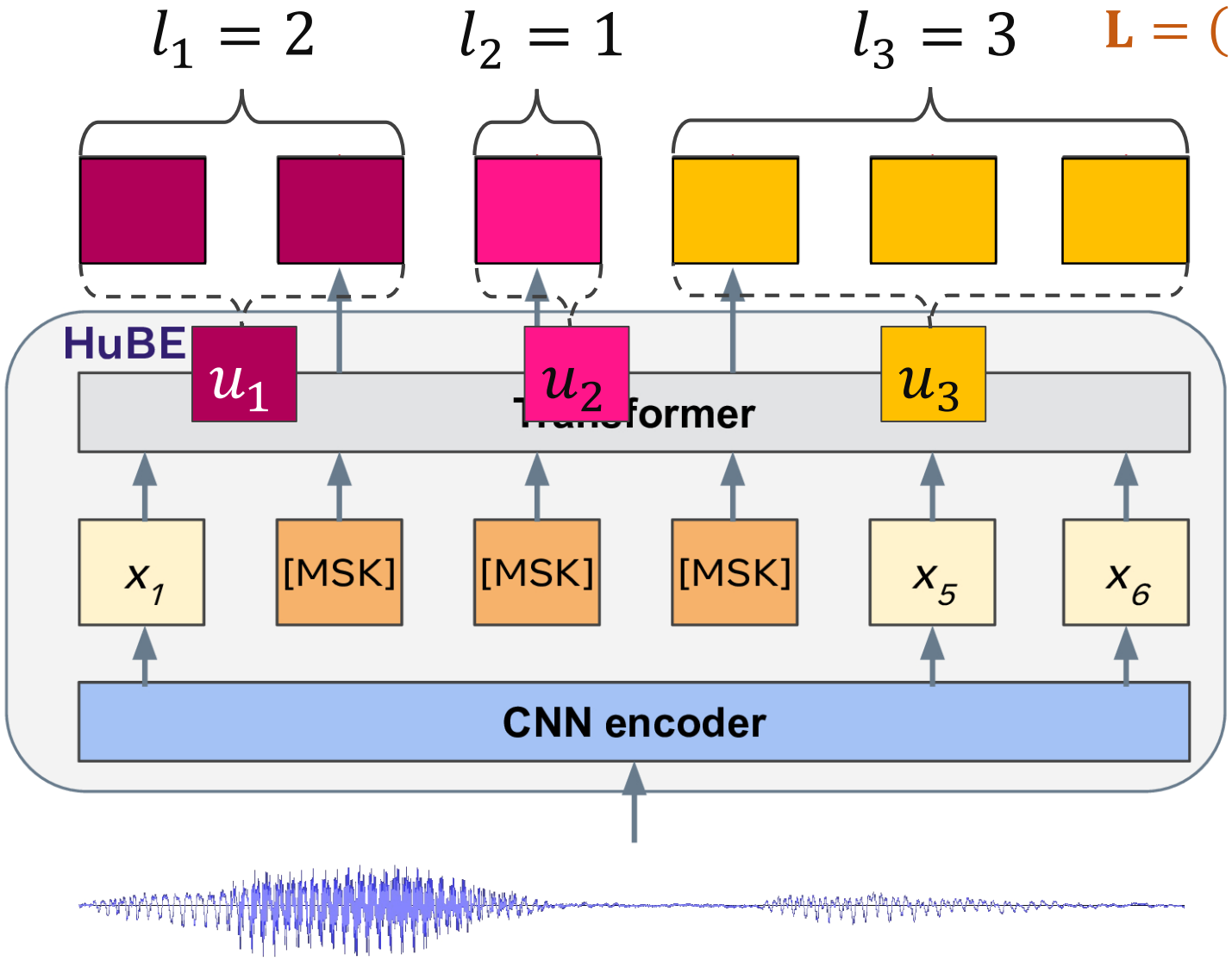
Adopt pretrained Wav2vec 2.0 as attribute encoders.

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

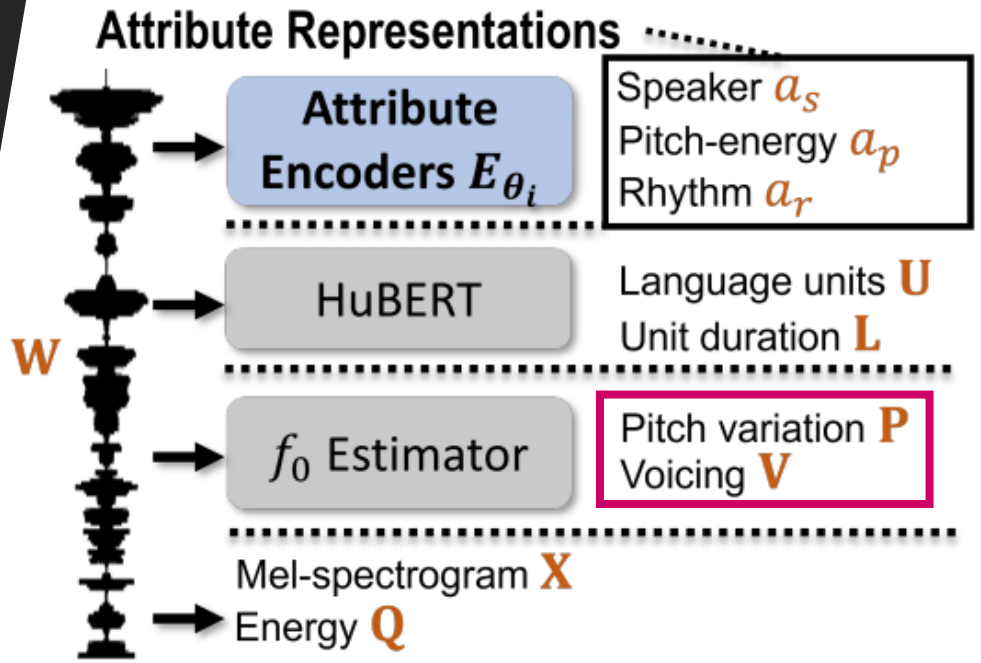
HuBERT


Acoustic unit discovery system
(e.g., K-means on MFCC)
z_1
z_2
z_3
z_4
z_5
z_6
HuBERT
Transformer
x_1
[MSK]
[MSK]
[MSK]
x_5
x_6
CNN encoder

HuBERT


$l_1 = 2$
$l_2 = 1$
$l_3 = 3$
$\mathbf{L} = (l_k \in \mathbb{N})_{k=1}^{K}$
$\sum_k l_k = J$
HuBE
u_1
u_2
u_3
Transformer
$\mathbf{U} = (u_k \in \mathcal{U})_{k=1}^{K}$
x_1
[MSK]
[MSK]
[MSK]
x_5
x_6
CNN encoder

Pitch, Voicing & Energy

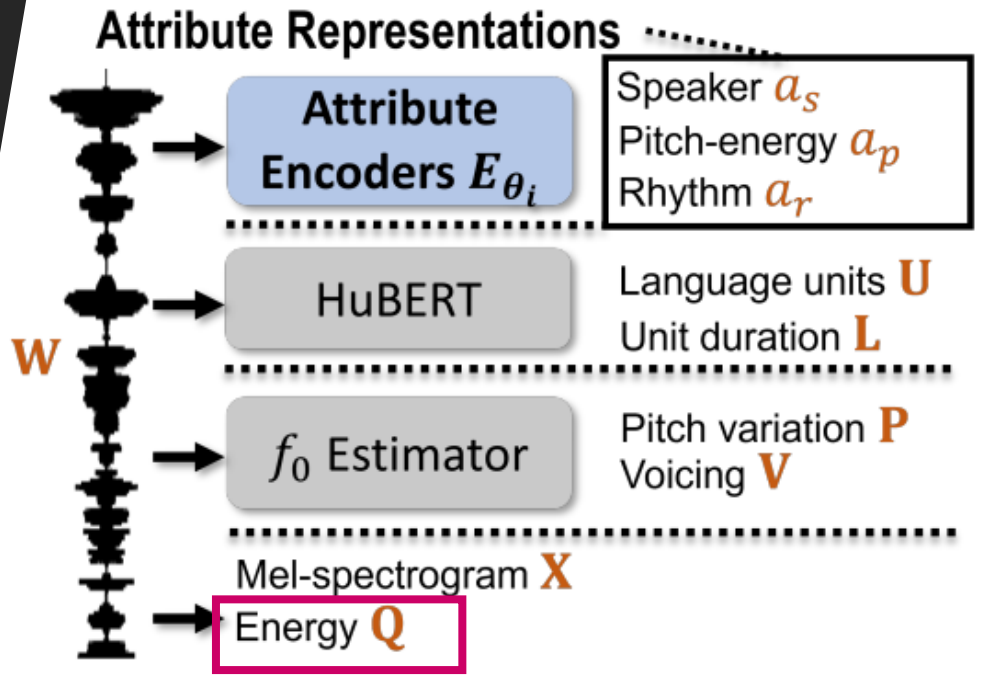


$\mathbf{P} = (p_j \in \mathbb{R})_{j=1}^{J}$ is mean-normalized pitch in Hertz.

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

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

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.

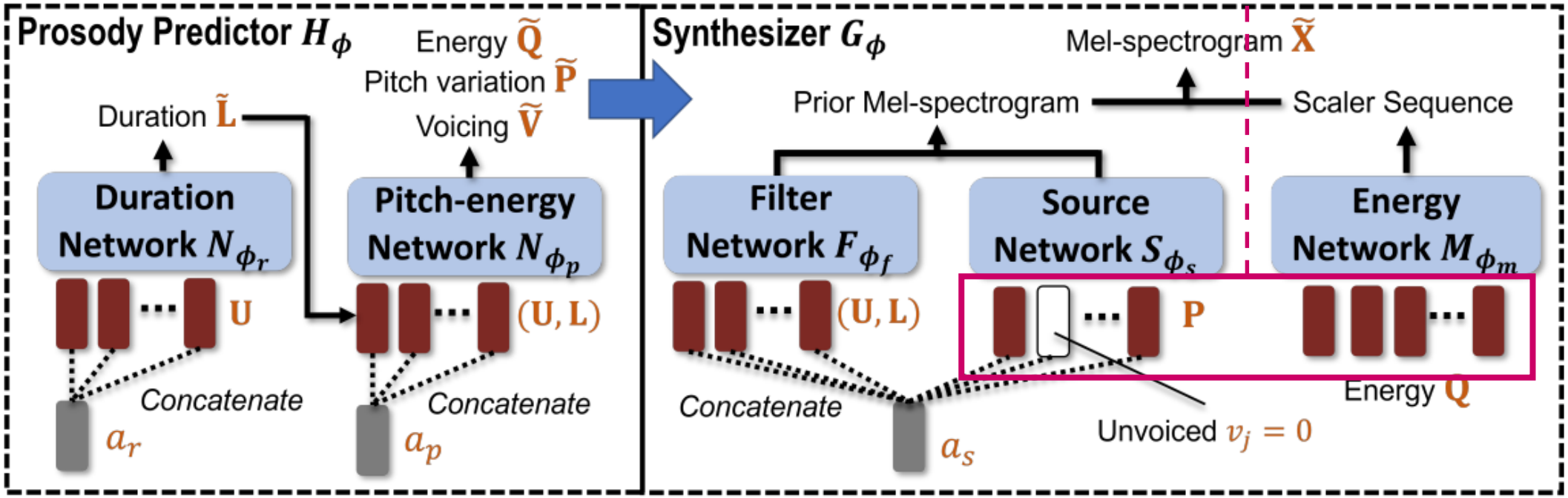
Obtain $\mathbf{P}$ and $\mathbf{V}$ from pitch estimator such as CREPE, YAPPT.

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

Convert $\mathbf{P}, \mathbf{Q}$ to representation.

1. Scalar to bin weight

$$b(h_j) = \{b_i(h_j)\}_{i=1}^{B},\ b_i(h_j) = exp\left(-\frac{(h_j - c_i)^2}{2\sigma^2}\right)$$
$$c_i = i\ell_b + p_{min}, h \in \{p, q\}$$

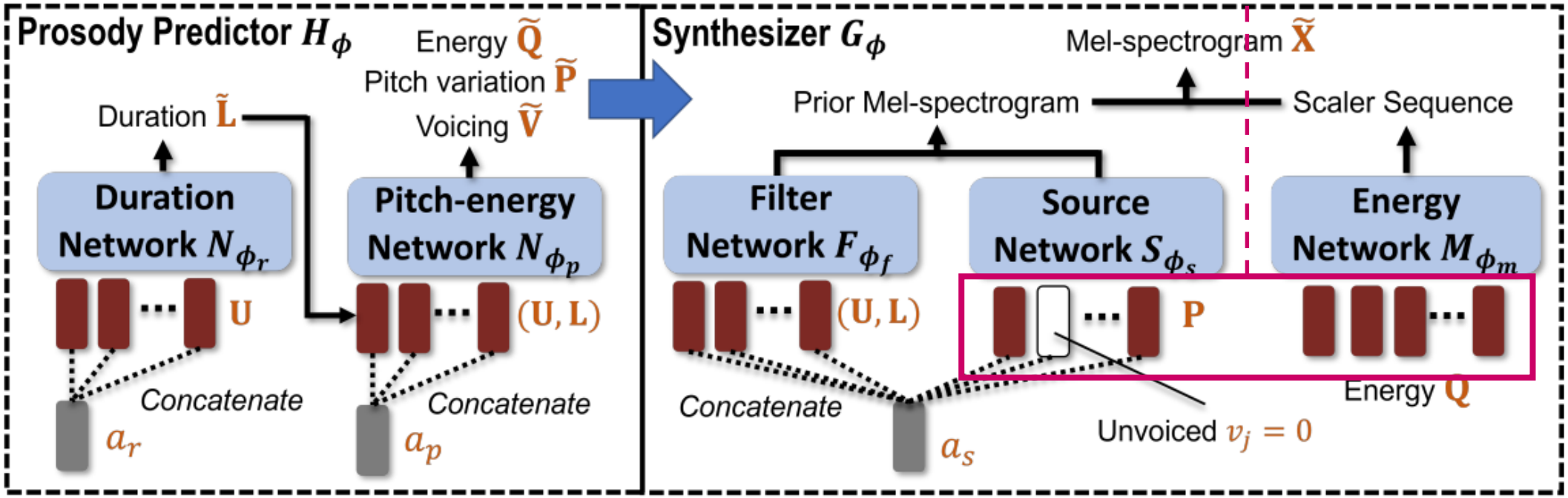


Convert $\mathbf{P}, \mathbf{Q}$ to representation.

2. Weighted sum with learnable embedding

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

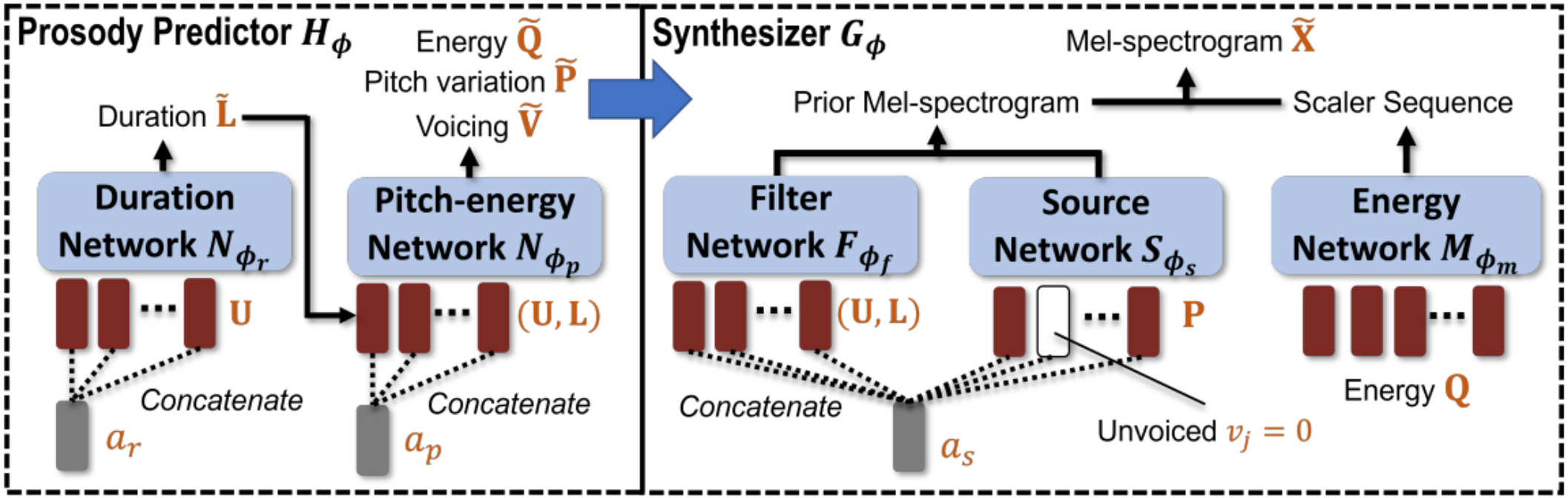
3. replace the unvoiced frames (j where $\mathrm{v}_j = 0$) of with $o(p_j, E^p)$ another learnable embedding.



Joint Optimization

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

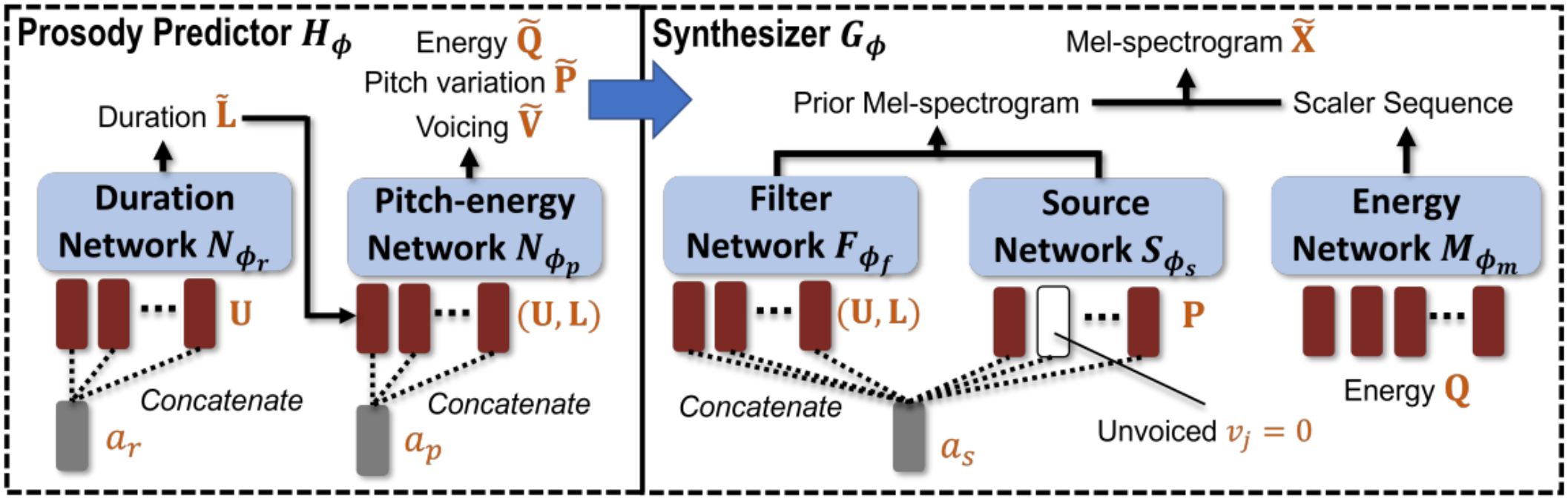
the poor estimation in the early stage of training is not conducive to G_ϕ considering prosody information, which will cause G_ϕ to ignore prosody information.



Joint Optimization

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

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

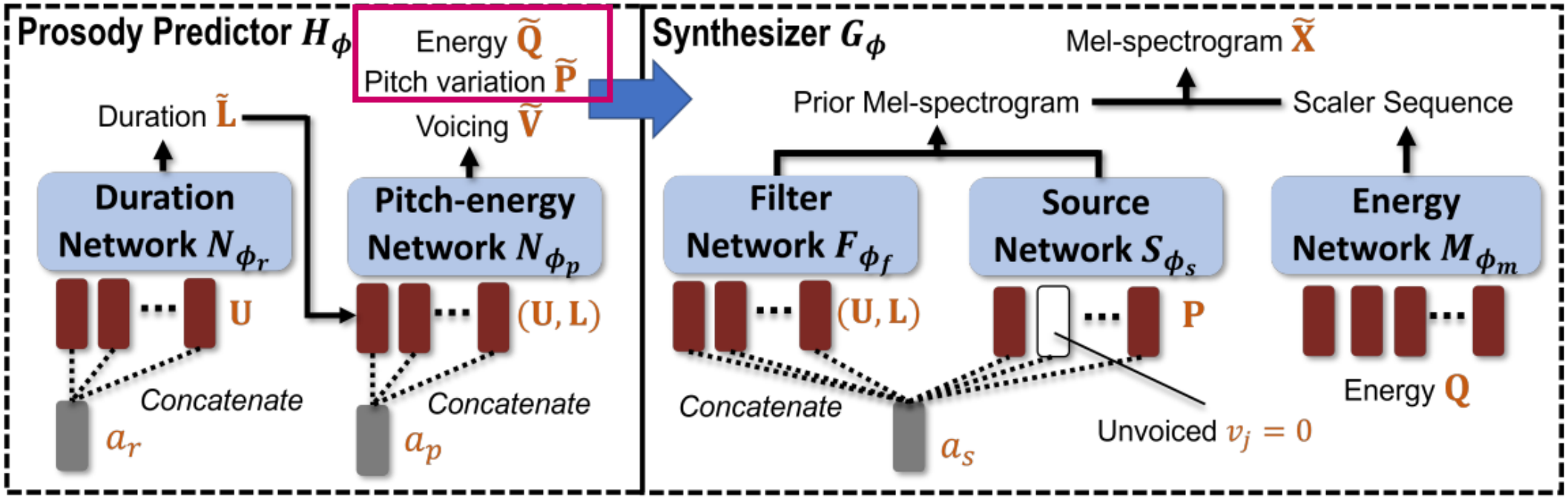


Objective Function

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

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

Predict bin weight

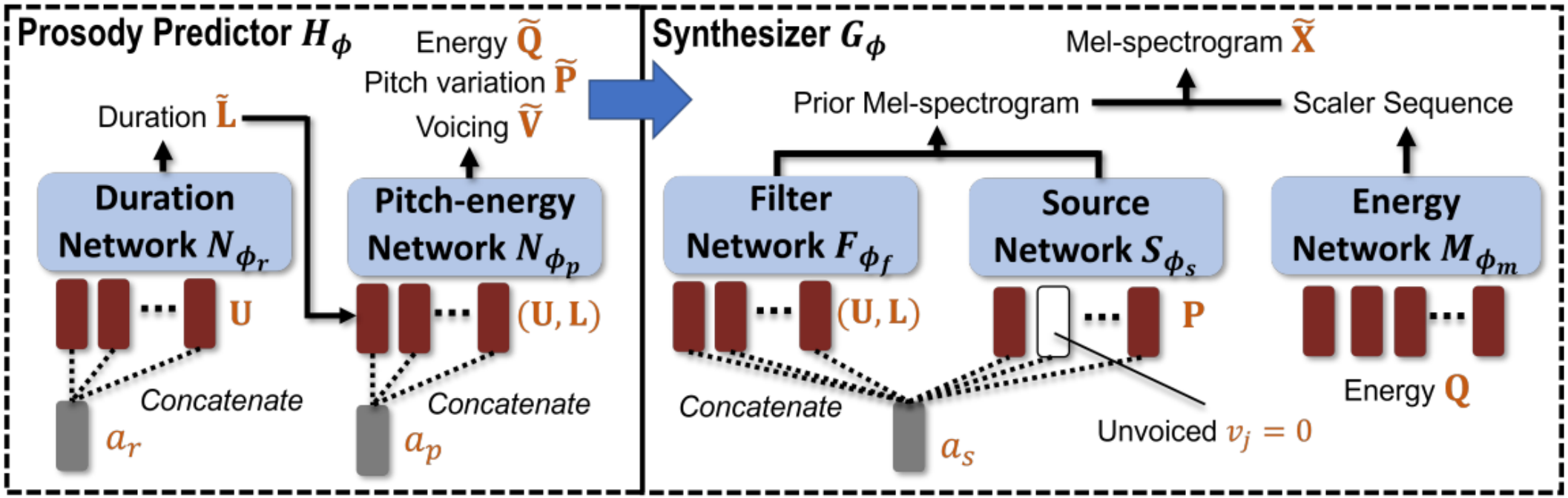


Objective Function

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

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

L1 loss

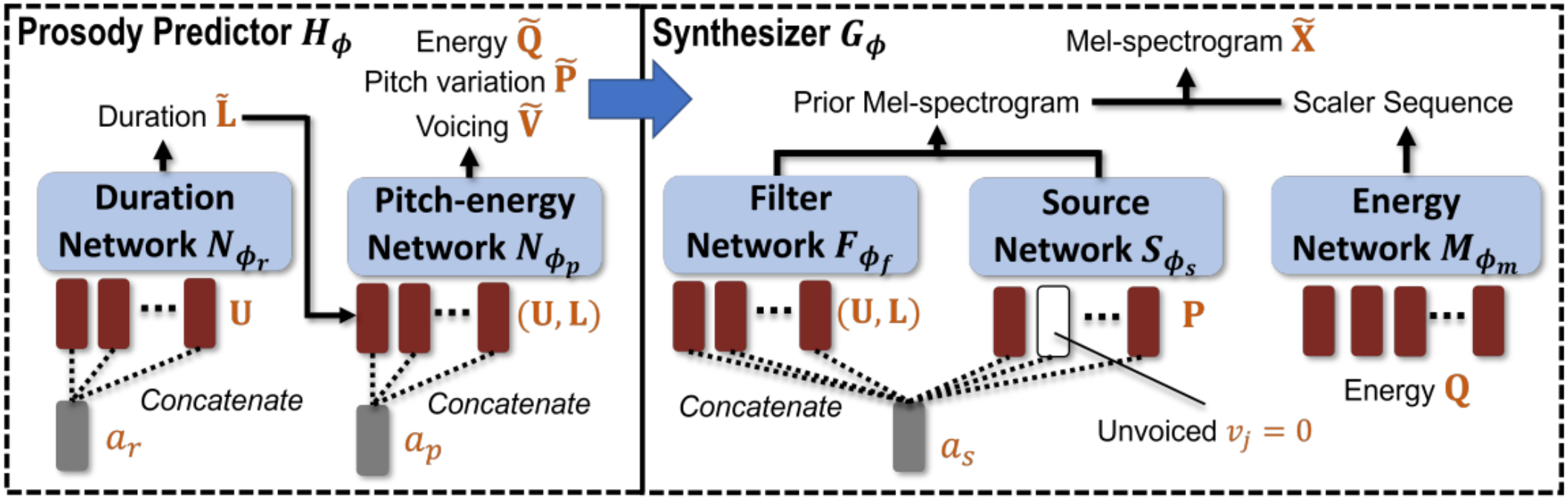


Objective Function

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

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

BCE loss of bin &
MSE loss of representation

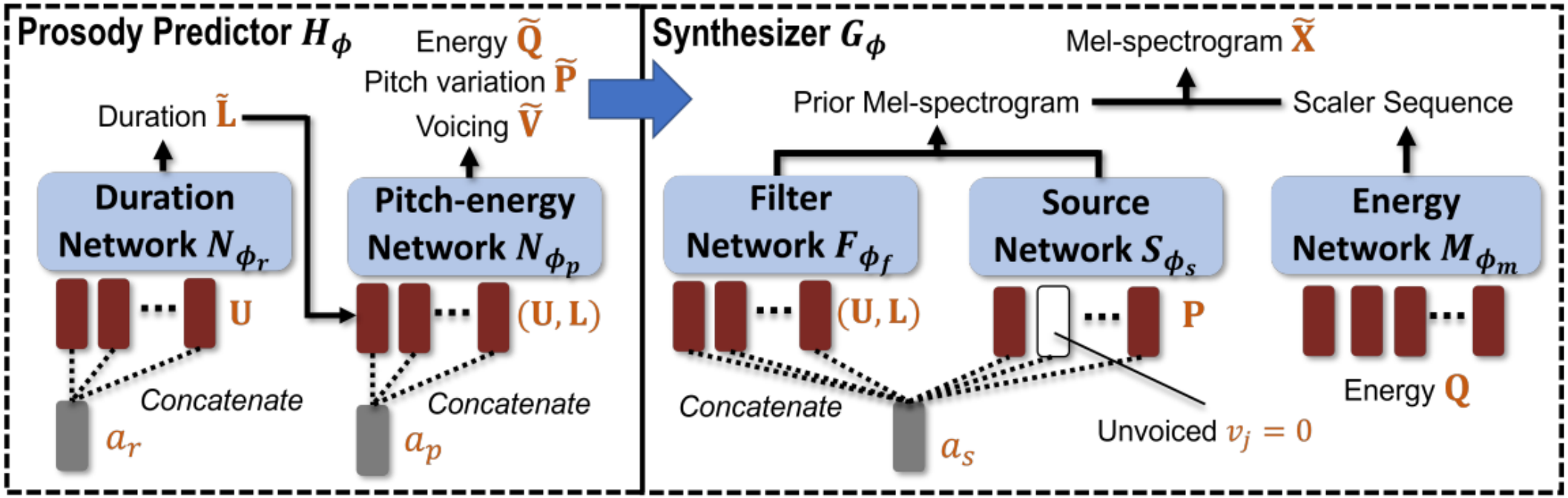


Objective Function

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

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

BCE loss

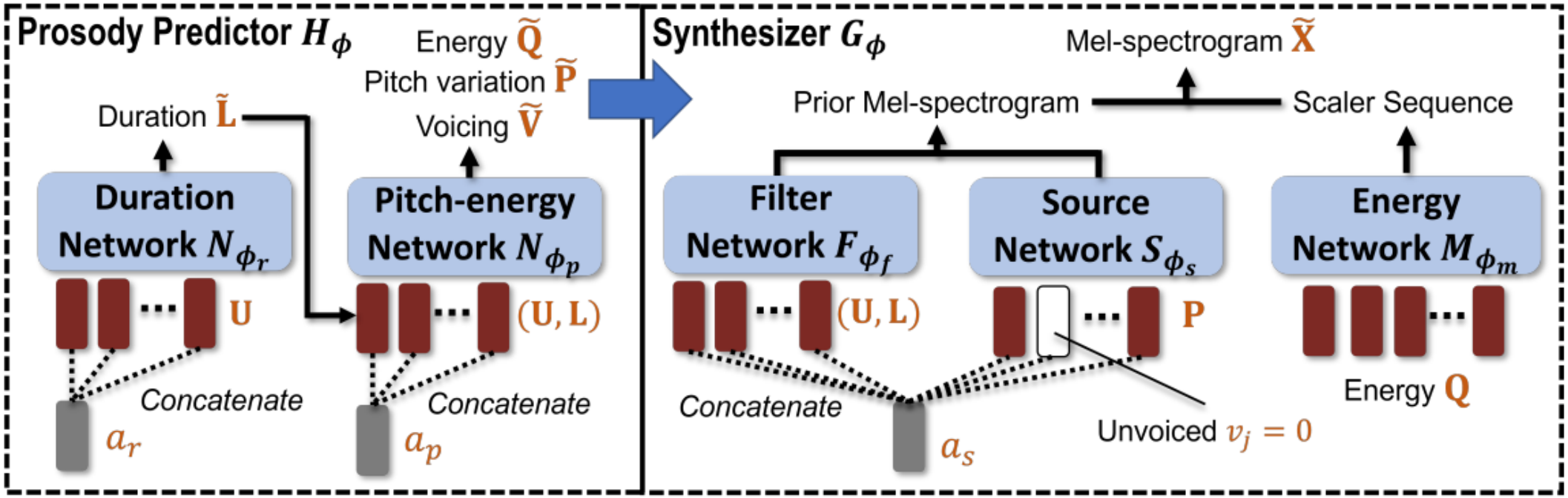


Objective Function

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

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

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 ↑
GT	5.5%	*n/a*	3.95 ± 0.11	*n/a*
AutoVC	88.4%	0.14	2.58 ± 0.17	2.62 ± 0.15
SRDVC	34.7%	0.17	3.25 ± 0.15	2.56 ± 0.14
Ours	7.5%	**0.34**	3.50 ± 0.14	$\mathbf{2.80} \pm 0.14$
(-W2V2)	7.8%	0.27	3.47 ± 0.13	2.62 ± 0.15
(-Joint Opt.)	**7.3%**	0.32	$\mathbf{3.55} \pm 0.14$	2.68 ± 0.14

SES: Speaker Embedding Similarity

Disentanglement of speech attributes

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

PCC	*AutoVC*	*SRDVC*	*Ours* $(\tilde{\mathbf{P}}, \tilde{\mathbf{Q}})$	*Ours* $(\mathbf{P}, \mathbf{Q})$
$\log f_0$	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 ↑
GT	5.5%	*n/a*	3.95 ± 0.11	*n/a*
SRDVC	49.9%	0.28	3.06 ± 0.16	2.58 ± 0.15
Ours	**8.9%**	**0.42**	$\mathbf{3.49} \pm 0.13$	$\mathbf{2.69} \pm 0.14$
(-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

Prosody Representations

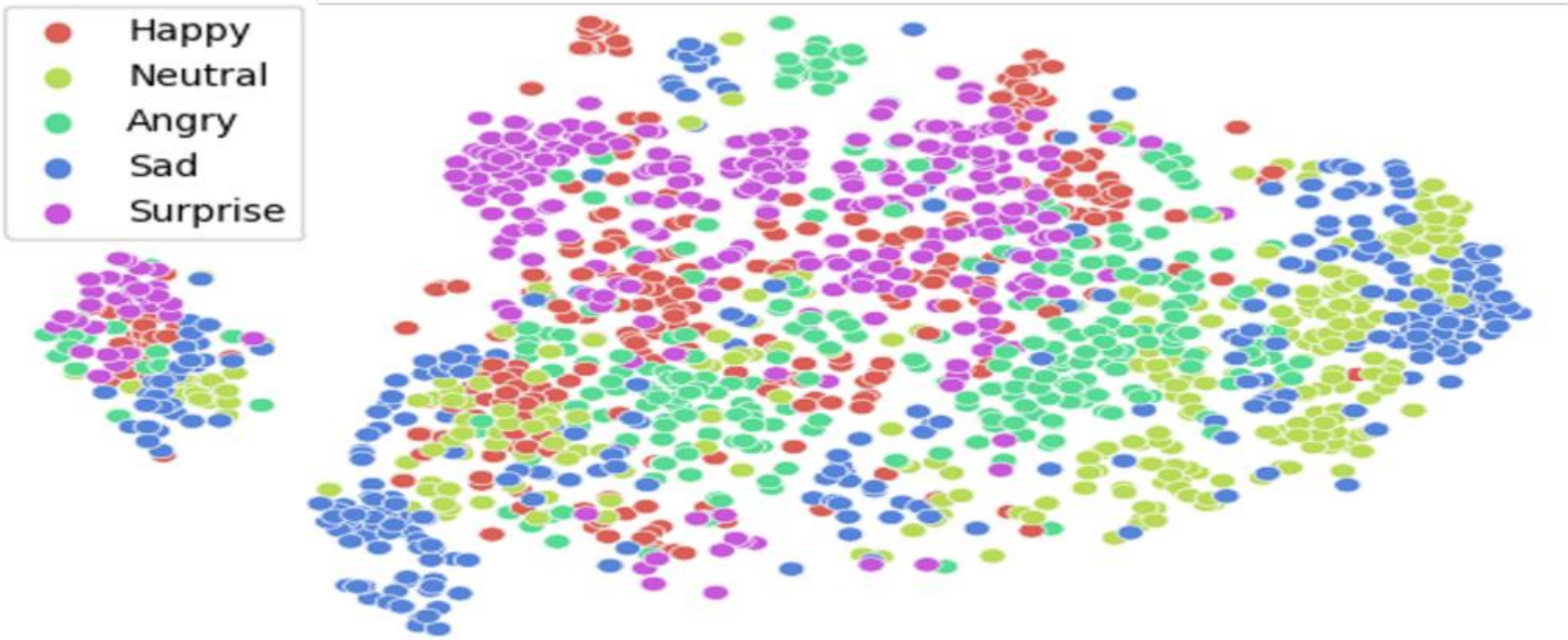


Fig. 2: Visualization of $\mathbf{a}_p$ (pitch-energy) by running t-SNE on ESD.

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 framewo
solve the problem
decomposition.

nt can be isolated
ext using HuBERT.

When cor
be decom

mation can