

Speech Processing: From Signal Processing to Deep Learning

Yossi Keshet

In memory of Yoav Medan



In memory of Yoav Medan



Super Resolution Pitch Determination of Speech Signals

Yoav Medan, Eyal Yair, and Dan Chazan

Abstract—Based on a new similarity model for the voice excitation process, a novel pitch determination procedure is derived. The unique features of the proposed algorithm are infinite (super) resolution, better accuracy than the difference limen for F_0 , robustness to noise, reliability, and modest computational complexity. The algorithm is instrumental to speech processing applications which require pitch synchronous spectral analysis.

I. INTRODUCTION

PITCH determination is considered one of the most difficult tasks in speech processing. Many pitch determination algorithms (PDA's) were proposed, both in the time and the frequency domains (e.g., [1]–[7]). The most comprehensive survey of PDA's is presented in [8], where it is claimed that “we do not have a single pitch determination algorithm which operates

sampling interval, contains a time quantization error which may lead to audible distortions in speech coding applications [9].

The PDA introduced in this paper overcomes most of these difficulties by introducing a new model for the similarity in the pitch process. This model allows quantifying the degree of similarity between exactly two adjacent and nonoverlapping pitch intervals, with an infinite time resolution. The similarity model takes into account the intensity modulation that may exist between successive periods to yield an instantaneous value of the pitch interval. The resulting algorithm offers a robust, high-resolution, and efficient implementation scheme which is capable of avoiding the audible distortions associated with common pitch based speech coding techniques. Given the high resolution and accuracy of the estimated pitch values afforded by the new approach, it was possible to develop an efficient and accurate

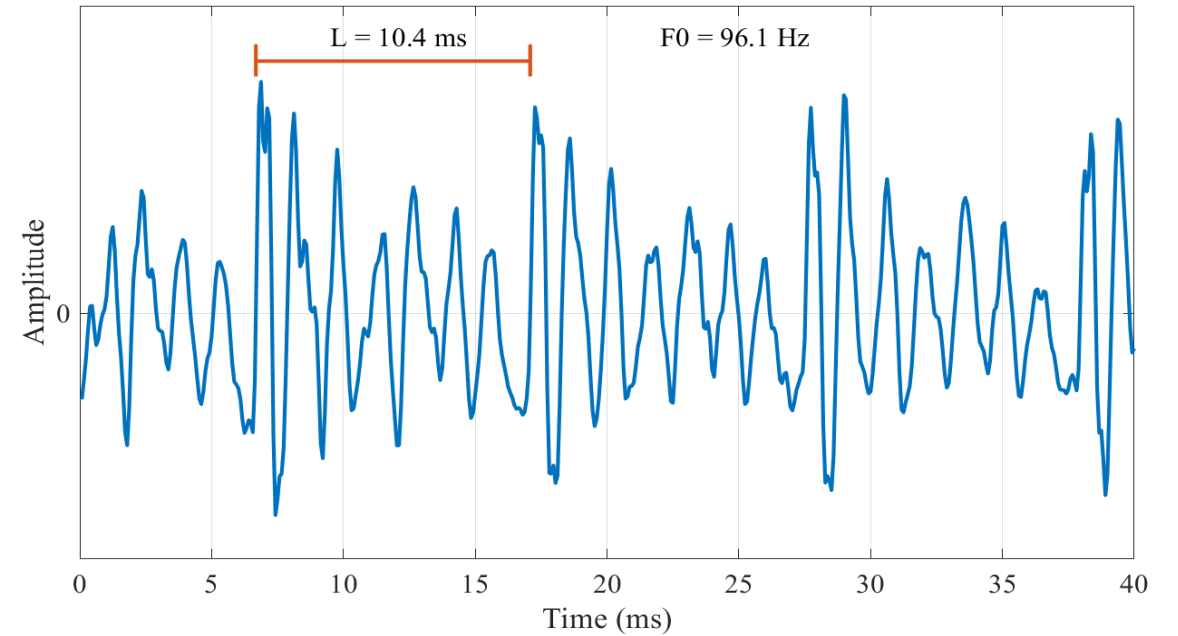
Outline

- Pitch estimation by multiple octave decoders
- Formant estimation and tracking using deep learning
- Speech time-scale modification by GANs
- Deep neural networks for speech steganography

Pitch estimation by multiple octave decoders

What is pitch, fundamental frequency or F0 ?

The fundamental frequency or F0 is the frequency at which vocal chords vibrate in voiced sounds.



Chuck Larson

When people hear their own voice through earphones, and when the voice pitch through the earphones is unexpectedly changed upwards or downwards, people automatically adjust the pitch of their voice.

This feedback mechanism does not work well for people with Autistic Syndrome Disorder (ASD)



Pitch estimation: goals

We propose a new model for pitch estimation that will have a better generalization by

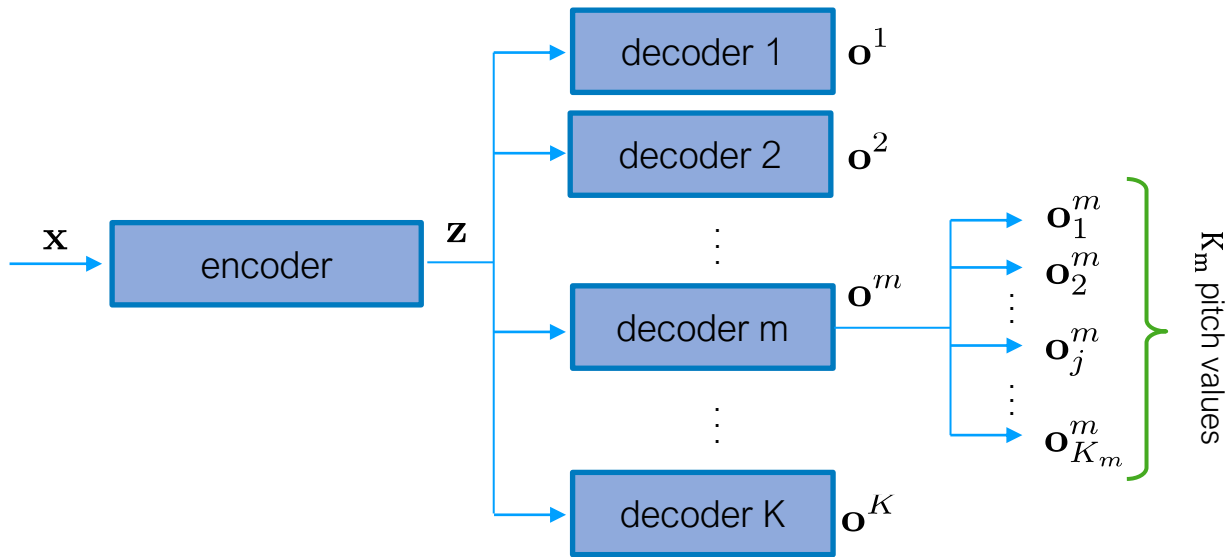
- signal representation suitable for pitch
- dynamic filters that are sufficiently narrow to resolve the harmonic at the pitch frequency and are sufficiently wide to integrate higher-order harmonics.

Pitch estimation: our approach

We propose a new model for pitch estimation that will have a better generalization by

- **An encoder** that learns a representation of the raw signal.
- **Multiple decoders**, where each decoder estimates:
 - Pitch value within a unique frequency band corresponds to a single octave
 - The confidence that the pitch is indeed in that frequency band.

Pitch estimation: our approach

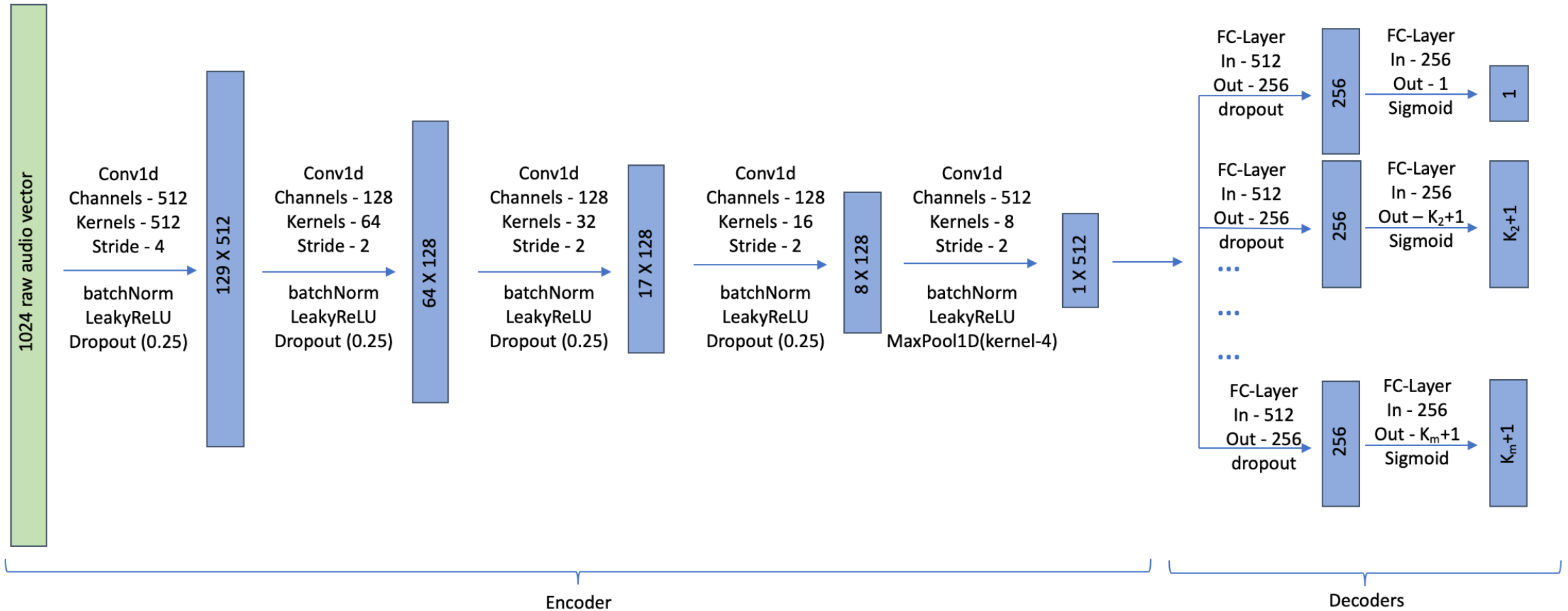


$$p^* = \operatorname{argmax}_{p,m} P(p|m, z)P(m|z)$$

$$\hat{P}(j|m, z) = \frac{\sigma(o_j^m)}{\sum_{i=0}^{K_m-1} \sigma(o_i^m)} \quad \hat{P}(m|z) = \sigma(o_{K_m}^m)$$

discrete pitch value j band m

Pitch estimation: our approach



Pitch estimation: our approach

We refer to our model as **PiMOD** (Pitch estimation by Multiple Octave Decoders).

COMPARISON OF PiMOD AND THE *BASELINE* ON THE MDB AND KEELE DATASETS. GREATER RPA AND VR VALUES INDICATES BETTER PERFORMANCE, WHEREAS LOWER GPE VALUES ARE BETTER

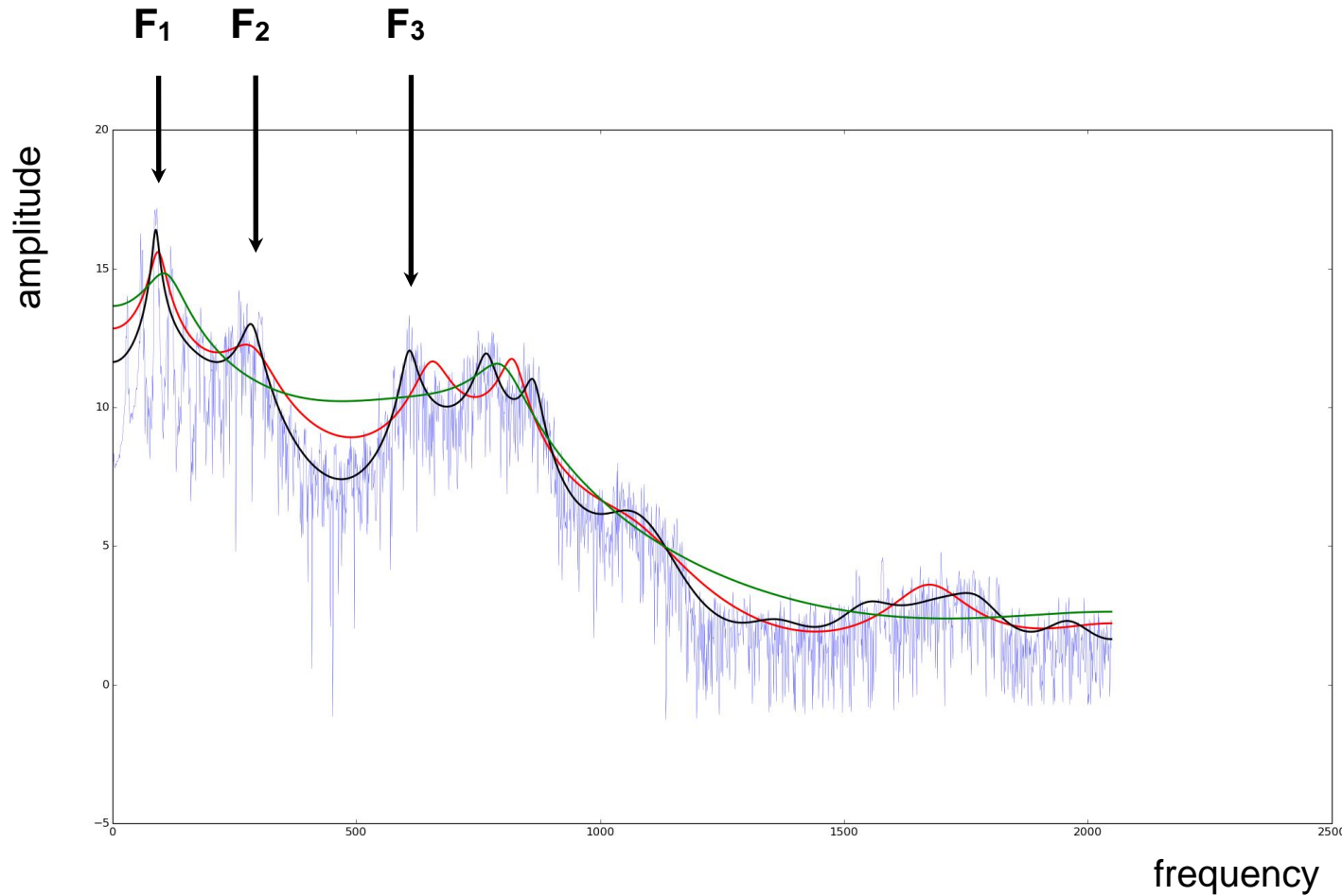
Dataset	Model	VR	RPA25	RPA50	GPE10
MDB	<i>baseline</i>	99.8±0.1	85.5±0.4	92.5±0.1	3.50±0.17
	PiMOD	99.9±0.00	86.0±0.3	93.4±0.1	2.46±0.14
Keele	<i>baseline</i>	95.8±0.7	72.7±2.0	86.6±1.0	7.18±0.41
	PiMOD	96.1±0.8	73.3±2.6	86.6±1.2	6.69±0.69

COMPARING PiMOD TO OTHER ALGORITHMS ON THE MDB AND KEELE DATASETS. GREATER RPA AND VR VALUES INDICATES BETTER PERFORMANCE, WHEREAS LOWER GPE VALUES ARE BETTER

Dataset	Model	VR	RPA25	RPA50	GPE10
MDB	PRAAT [13]	89.3	67.0	75.5	16.2
	pYIN [12]	89.0	44.0	58.5	19.4
	SWIPE [16]	84.7	70.7	77.3	18.1
	CREPE [19]	99.7±0.1	84.9±0.6	92.2±0.5	2.9±0.23
	CREPE-D	99.8±0.1	85.0±0.4	92.3±0.4	2.8±0.27
	PiMOD	99.9±0.00	86.0±0.3	93.4±0.1	2.46±0.14
Keele	PRAAT [13]	90.9	56.8	76.9	11.4
	pYIN [12]	89.6	49.8	71.4	12.1
	SWIPE [16]	83.1	52.3	73.9	17.7
	CREPE [19]	95.0±0.3	73.1±2.0	86.2±0.7	7.67±0.25
	CREPE-D	95.7±0.2	73.9±2.0	86.4±0.7	7.22±0.26
	PiMOD	96.1±0.8	73.3±2.6	86.6±1.2	6.69±0.69

Formant estimation and tracking using deep networks

What are formants?



Formants are considered to be resonances of the vocal tract during speech production.

The formant frequencies approximately correspond to the peaks of the spectrum of the vocal tract. These peaks cannot be easily extracted from the spectrum, since the spectrum is also tainted with pitch harmonics.

Does the Queen speak the Queen's English?



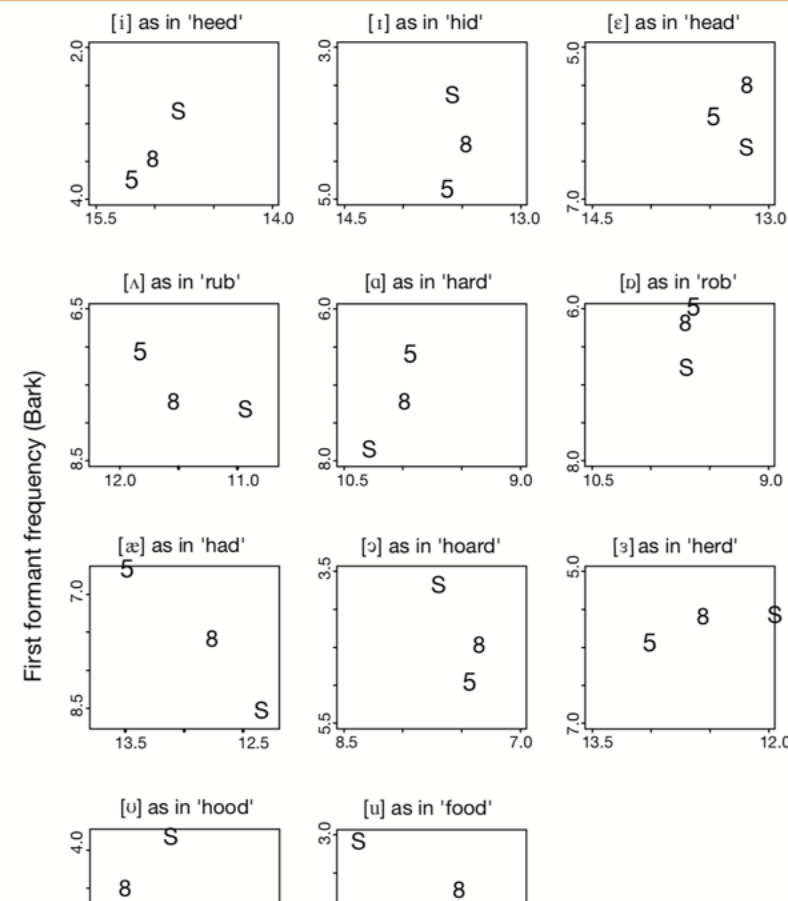
The Queen's

Does the Queen speak the Queen's English?

Elizabeth II's traditional pronunciation has been influenced by modern trends.

The pronunciation of all languages changes subtly over time¹, mainly owing to the younger members of the community². What is unknown is whether older members unwittingly adapt their accent towards community changes. Here we analyse vowel sounds from the annual Christmas messages broadcast by HRH Queen Elizabeth II during the period between the 1950s and 1980s. Our analysis reveals that the Queen's pronunciation of some vowels has been influenced by the standard southern-British accent of the 1980s which is more typically associated with speakers who are younger and lower in the social hierarchy.

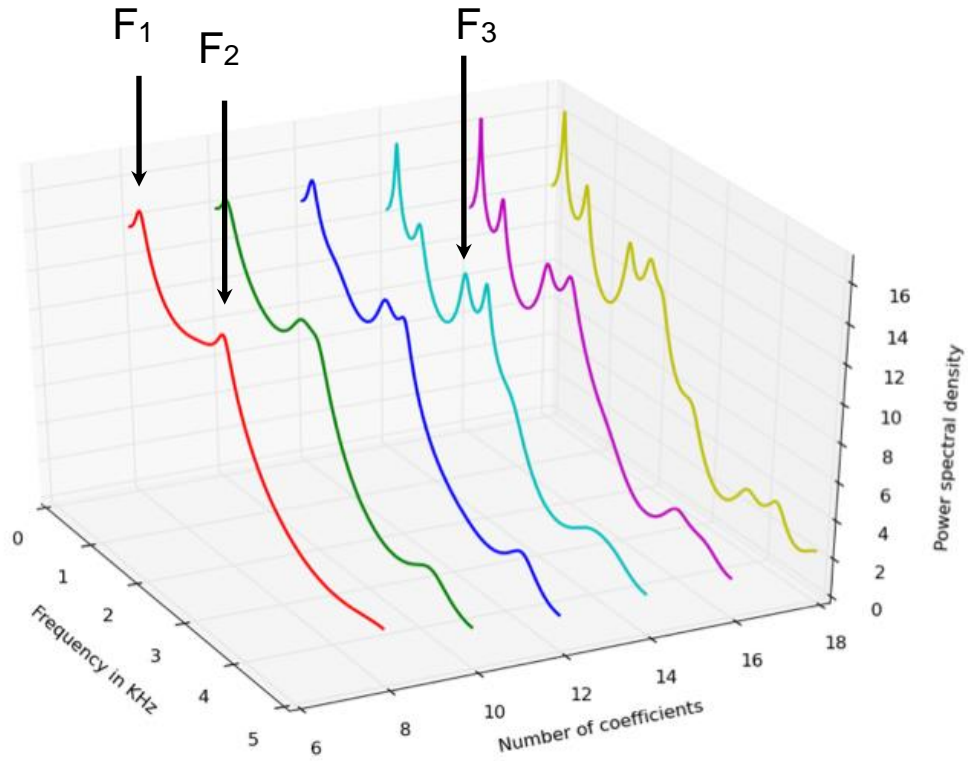
Phoneticians have documented many types of change to the standard accent of British English known as 'received pronunciation'³, some of which have a corollary in the changing attitudes towards social class. There was a marked social stratification in Britain in the 1950s⁴, and in 1963 the phonetician David Abercrombie wrote, "One either speaks received pronunciation, or one does not, and if the opportunity to learn it in youth has not arisen, it is almost impossible to learn it in later life"⁵. But as class distinctions have become more blurred⁴, so too have the linguistic distinctions between English



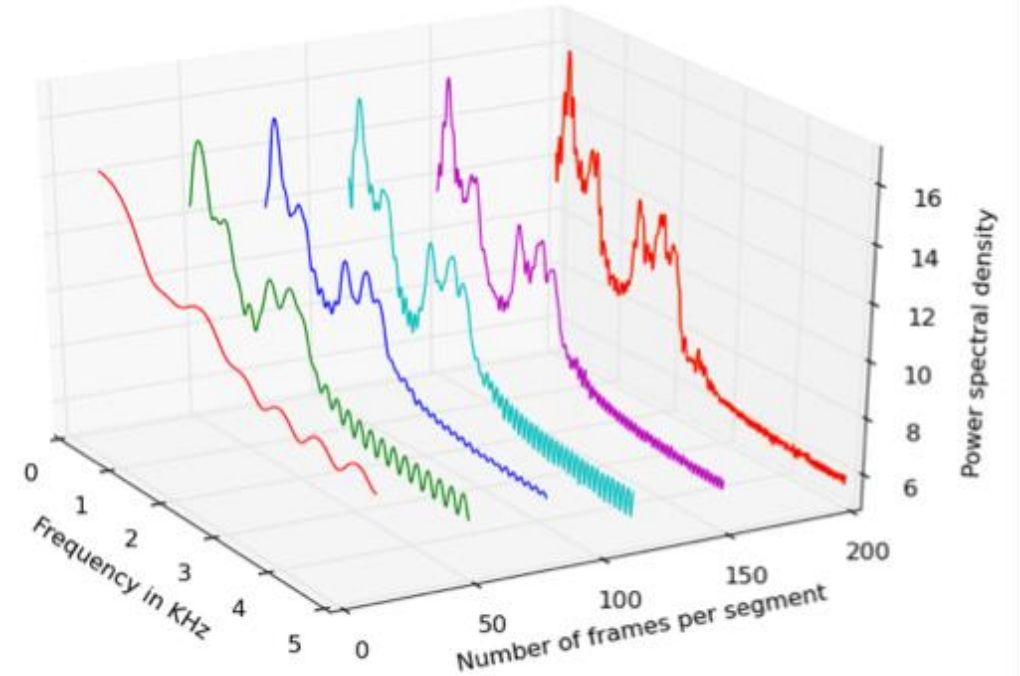
1985

ature, 2000

Formant estimation: new approach




Spectral envelop with different number of coefficients (Linear Predictive Coding;LPC)



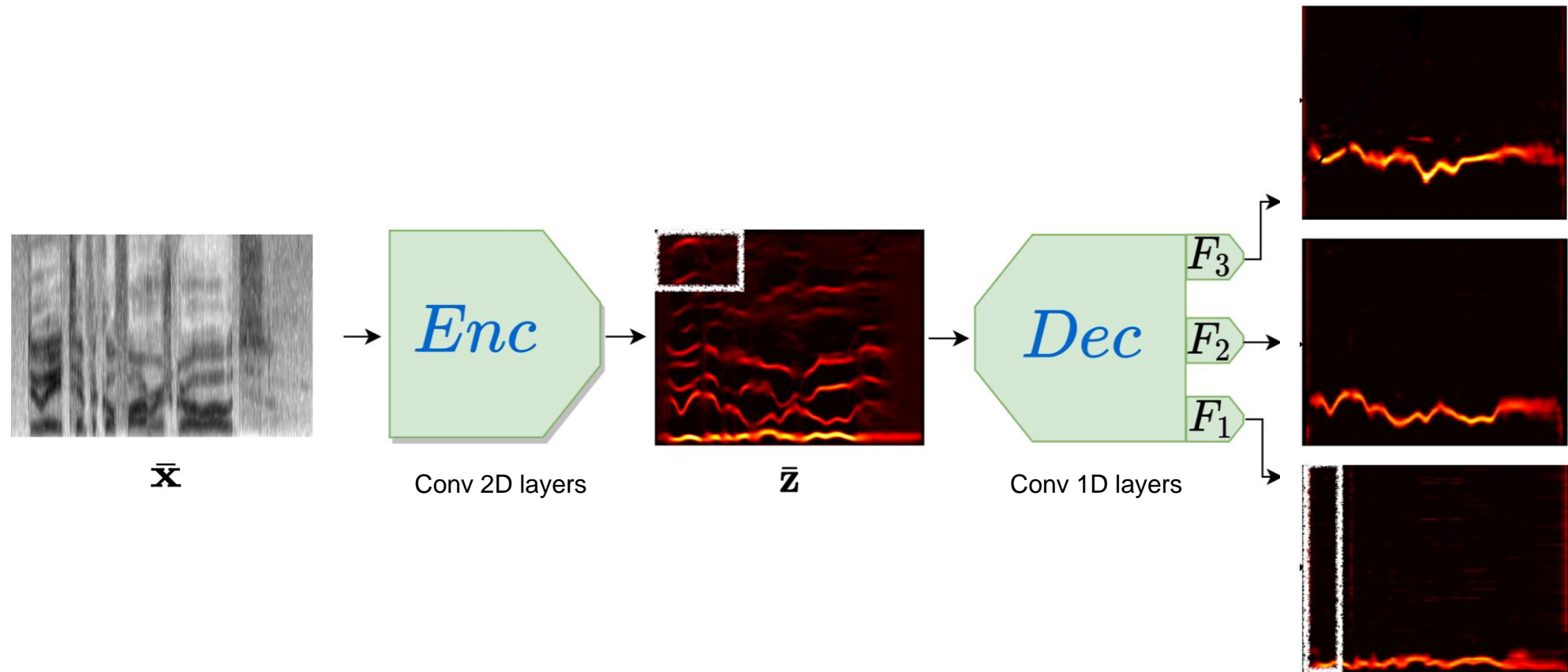
Pitch-synchronous spectrum, for different values of pitch (true pitch is 123 Hz)

Formant estimation: our approach I

	Error in Hz		
	F_1	F_2	F_3
Praat	130	230	267
WaveSurfer	70	94	154
MSR	64	105	125
DeepFormants	54	81	112
DeepFormants II (CNN)	45	65	94
inter-labler	55	69	84

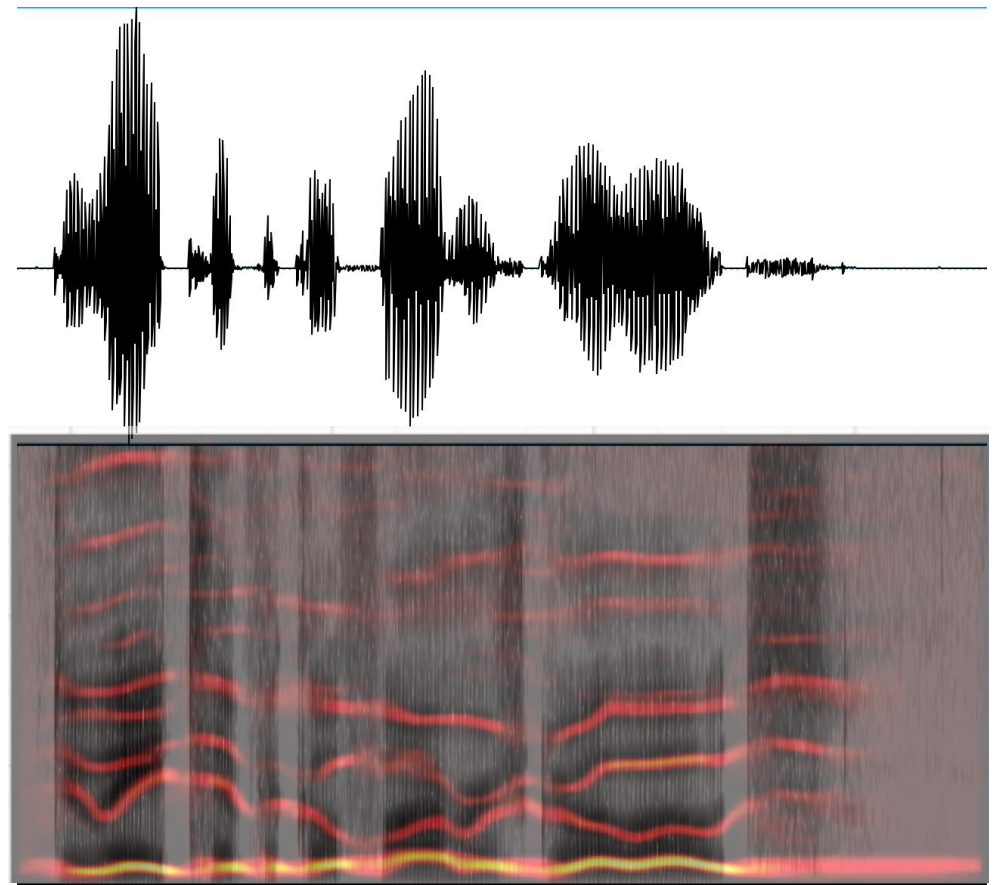


Formant estimation: our approach II



Formant estimation: our approach II

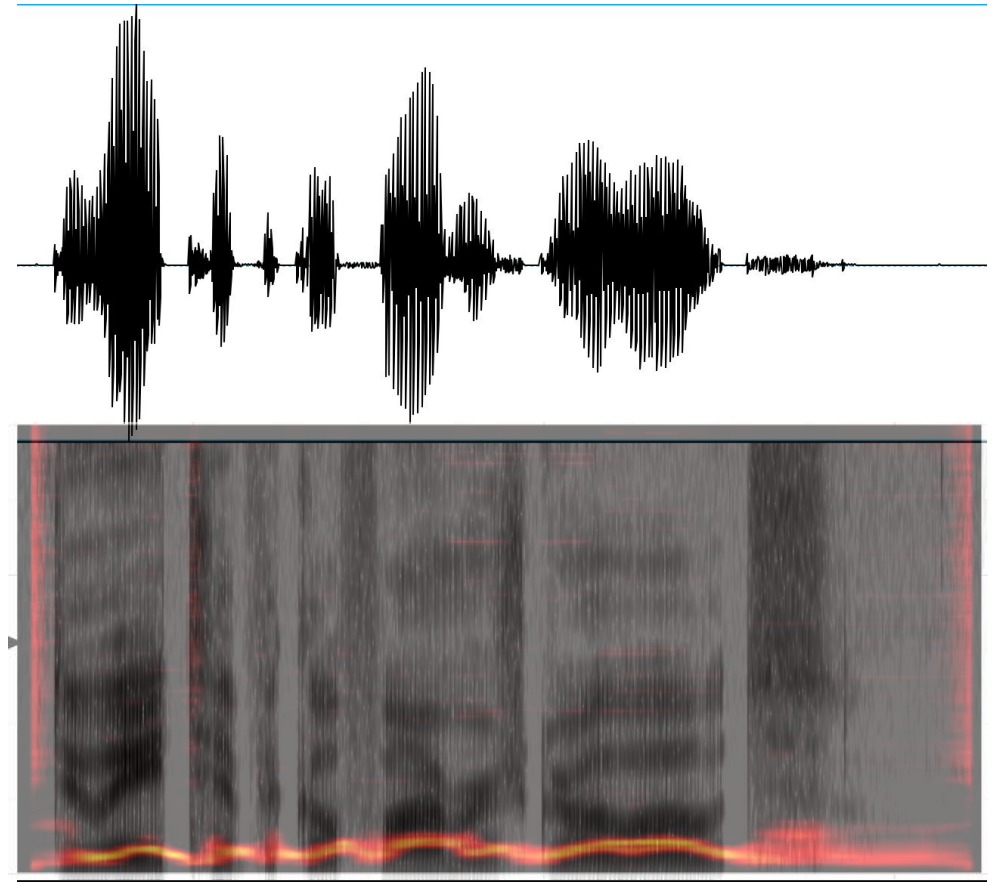
$$\bar{\mathbf{z}} = \text{Encoder}(\bar{\mathbf{x}})$$



Formant estimation: our approach II

$$\begin{aligned} & \Pr(F_1, F_2, F_3 | \bar{\mathbf{z}}) \\ &= \Pr(F_1 | \bar{\mathbf{z}}) \Pr(F_2 | F_1, \bar{\mathbf{z}}) \Pr(F_3 | F_2, F_1, \bar{\mathbf{z}}) \end{aligned}$$

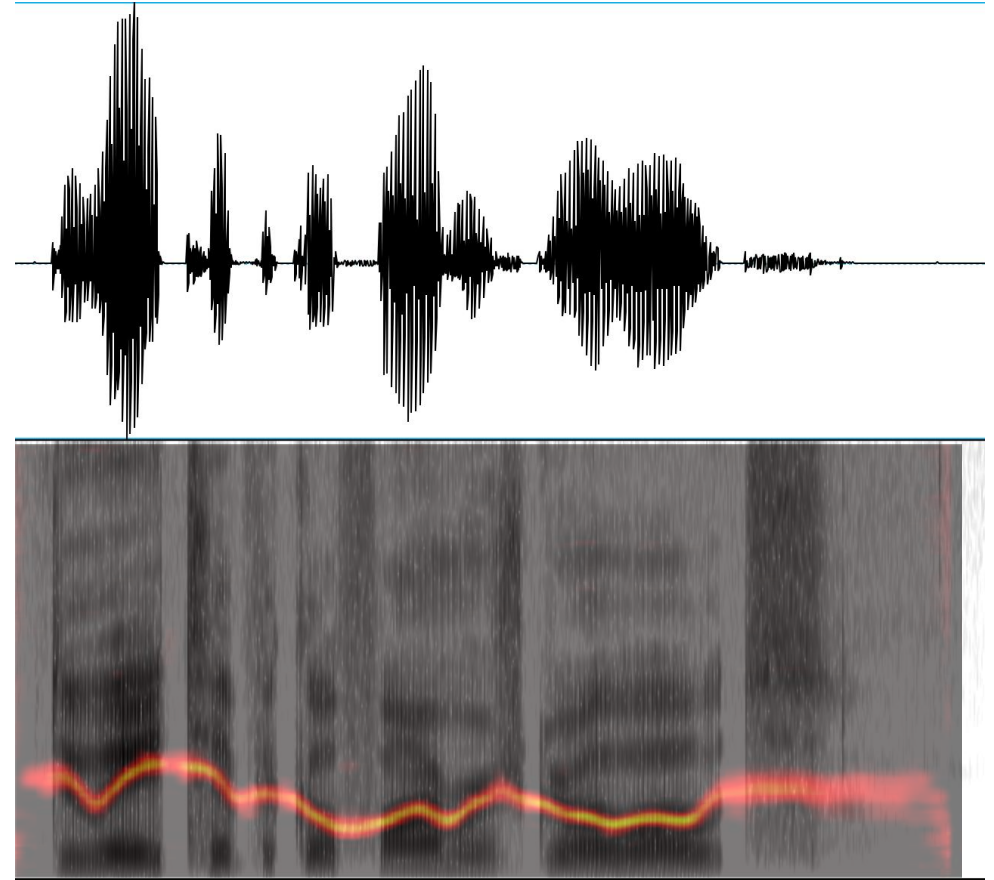
Decoder \mathbf{F}_1



Formant estimation: our approach II

$$\begin{aligned} \Pr(F_1, F_2, F_3 | \bar{\mathbf{z}}) \\ = \Pr(F_1 | \bar{\mathbf{z}}) \Pr(F_2 | F_1, \bar{\mathbf{z}}) \Pr(F_3 | F_2, F_1, \bar{\mathbf{z}}) \end{aligned}$$

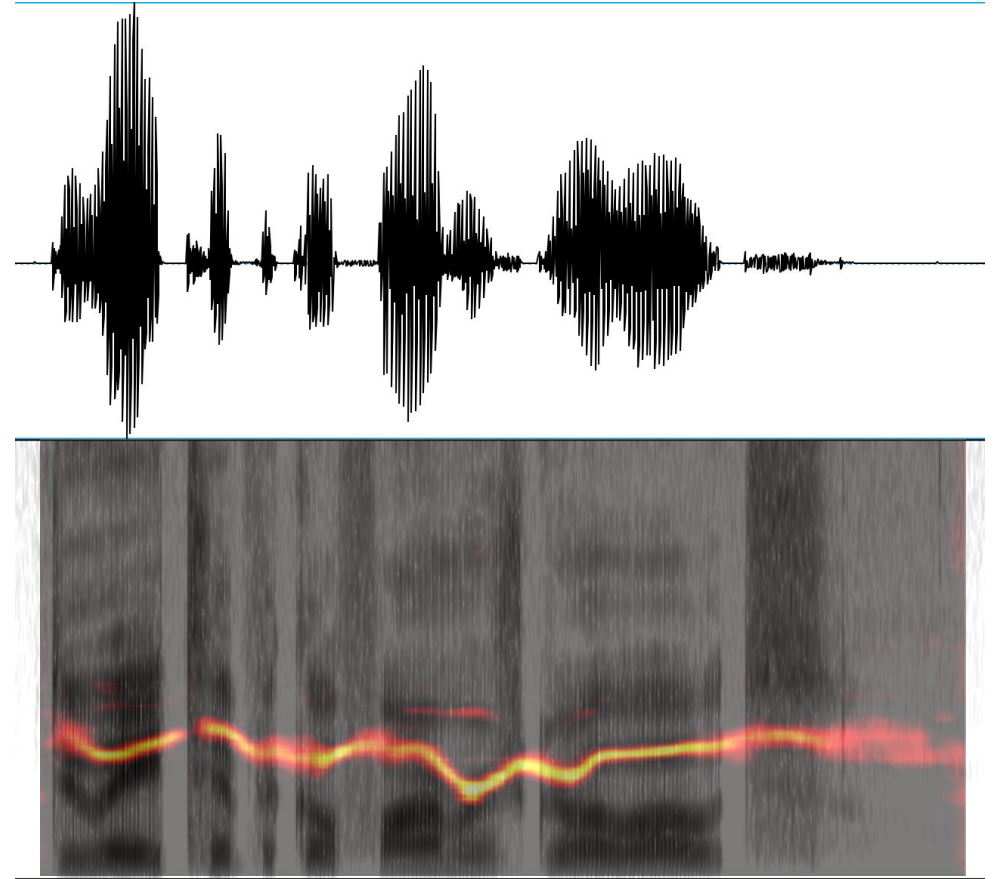
Decoder F_2



Formant estimation: our approach II

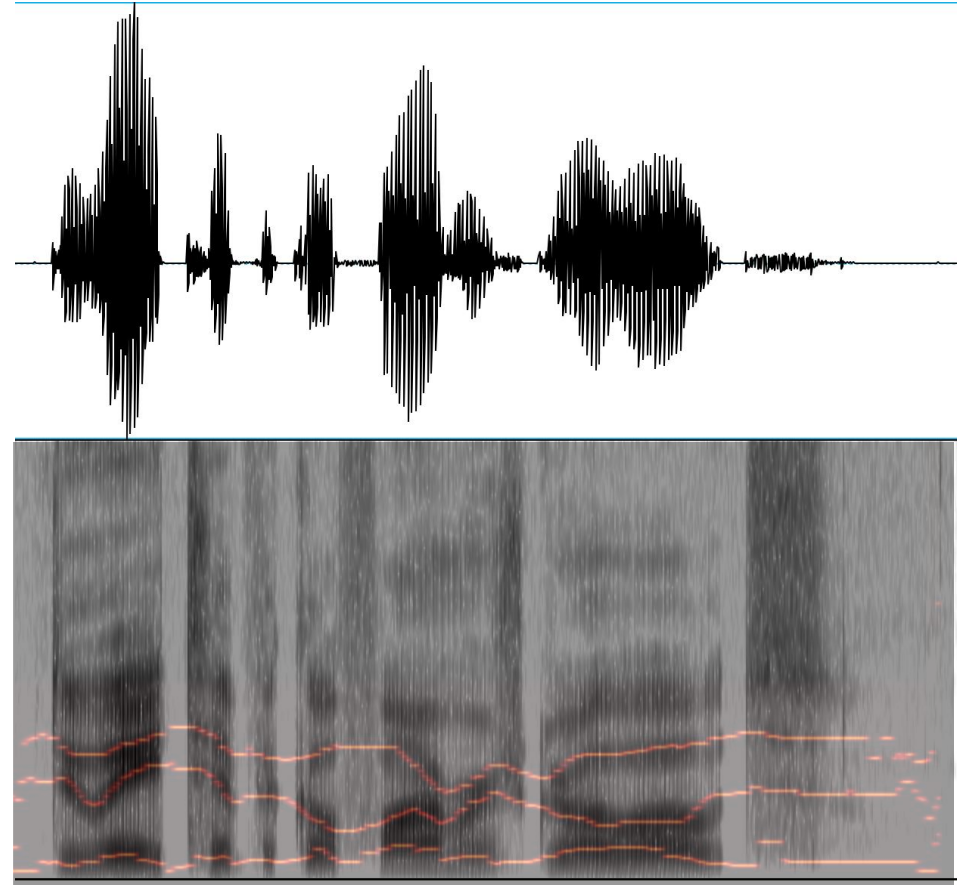
$$\begin{aligned} \Pr(F_1, F_2, F_3 | \bar{\mathbf{z}}) \\ = \Pr(F_1 | \bar{\mathbf{z}}) \Pr(F_2 | F_1, \bar{\mathbf{z}}) \Pr(F_3 | F_2, F_1, \bar{\mathbf{z}}) \end{aligned}$$

Decoder F_3



Formant estimation: our approach II

$$\begin{aligned} \Pr(F_1, F_2, F_3 | \bar{\mathbf{z}}) \\ = \Pr(F_1 | \bar{\mathbf{z}}) \Pr(F_2 | F_1, \bar{\mathbf{z}}) \Pr(F_3 | F_2, F_1, \bar{\mathbf{z}}) \end{aligned}$$



Formant estimation: our approach II

Dataset	Method	F_1	F_2	F_3
VTR	WaveSurfer	70	96	154
	DeepFormants	50	86	104
	Ours	39	30	47
Hillenbrand	WaveSurfer	68	190	182
	DeepFormants (Train VTR)	71	160	131
	DeepFormants (Train All)	36	100	116
	Ours (Train VTR)	74	150	125
	Ours (Train All)	26	78	82
Clopper	WaveSurfer	128	181	—
	DeepFormants (Train VTR)	228	168	—
	DeepFormants (Train All)	103	157	—
	Ours (Train VTR)	99	147	—
	Ours (Train All)	49	64	—

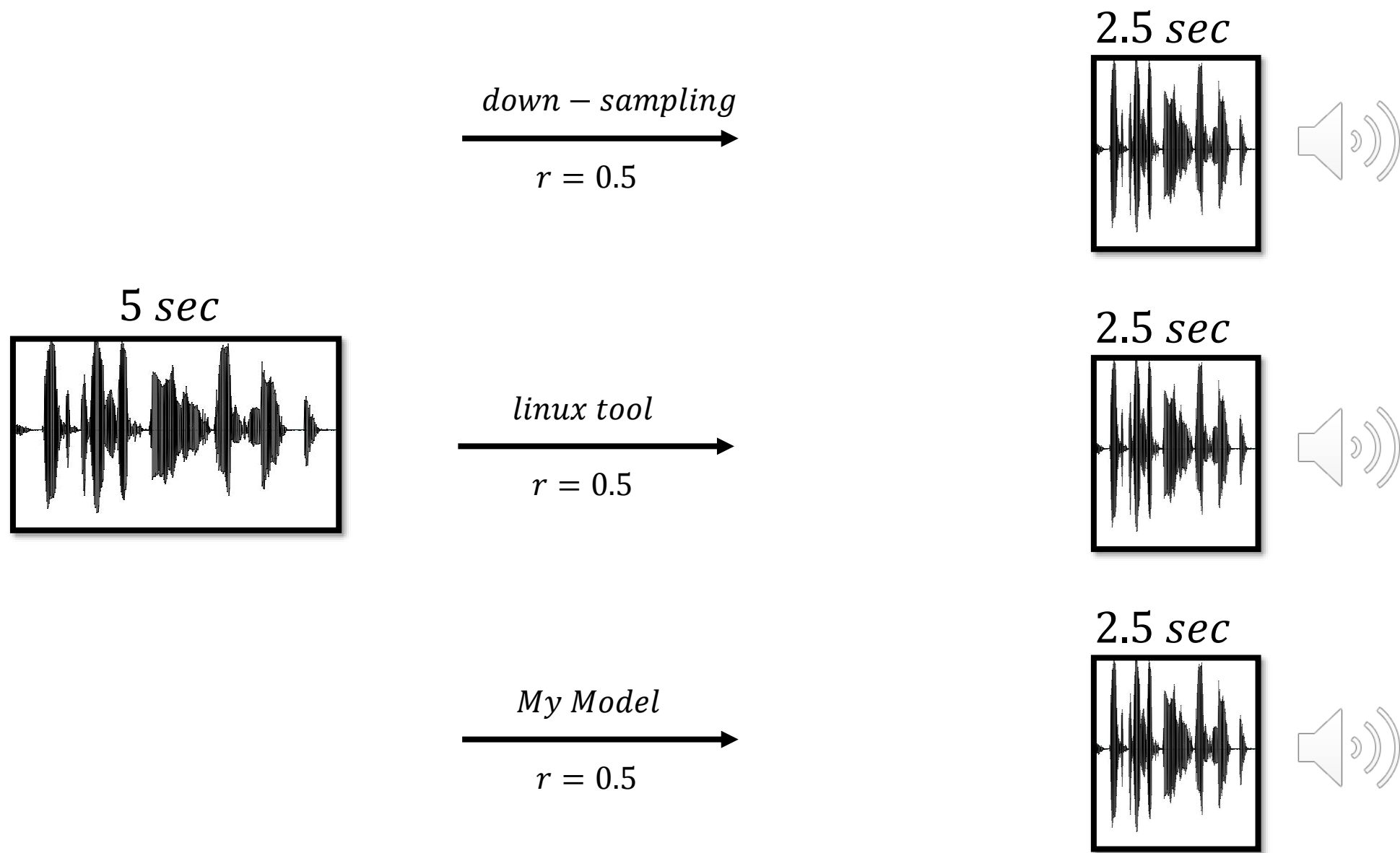
Time-scale modification of speech

Time-scale modification of speech

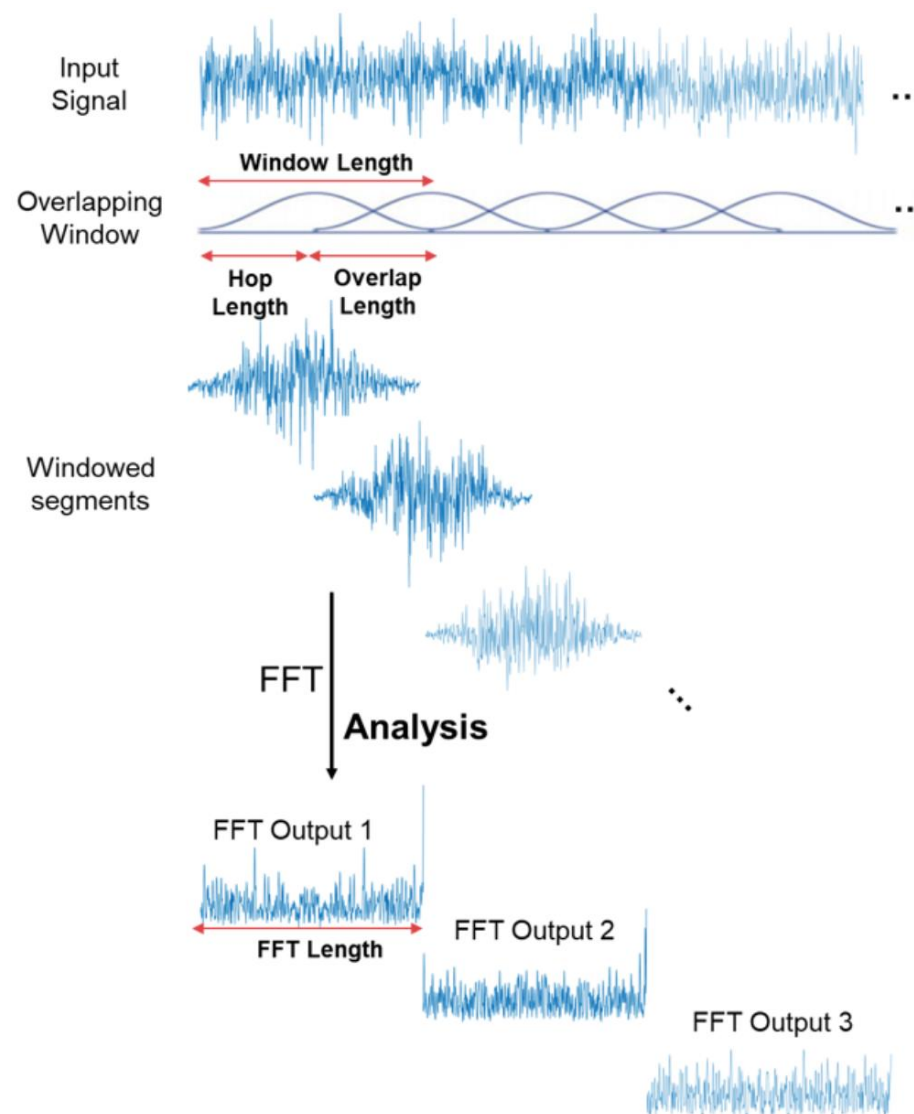
When you want to speed-up or slow-down the speech...



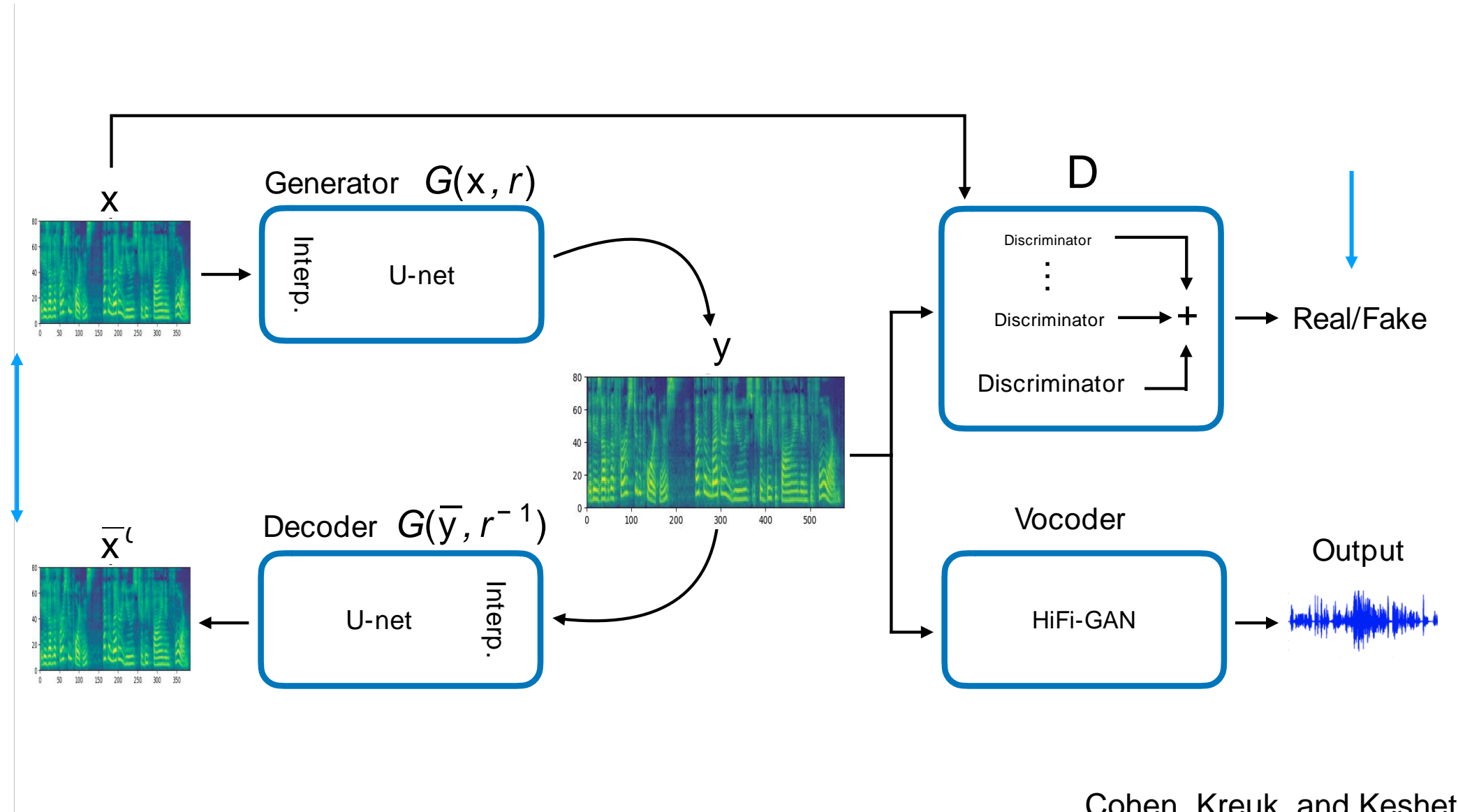
Apple
Podcasts



Time-scale modification of speech: signal processing

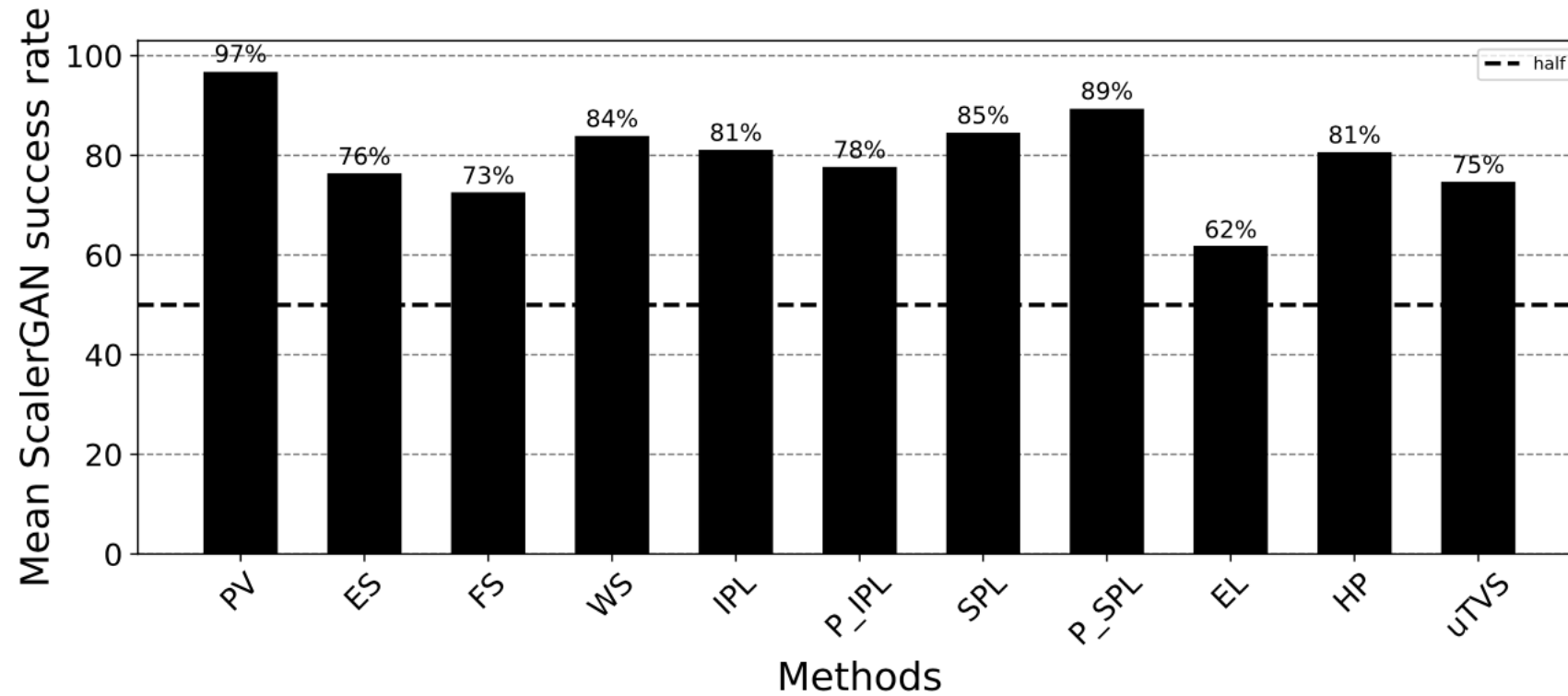


Time-scale modification of speech: deep learning



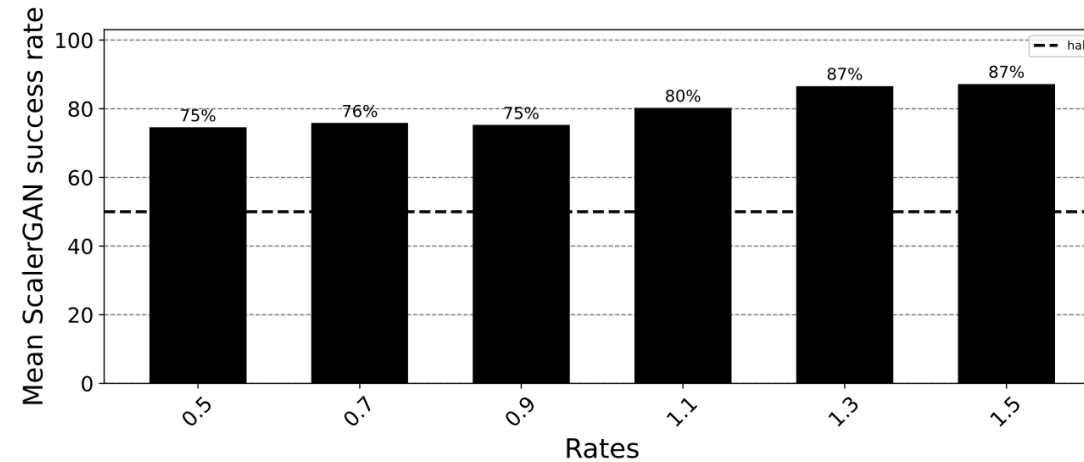
Time-scale modification of speech: deep learning

Aggregation by method



Time-scale modification of speech: deep learning

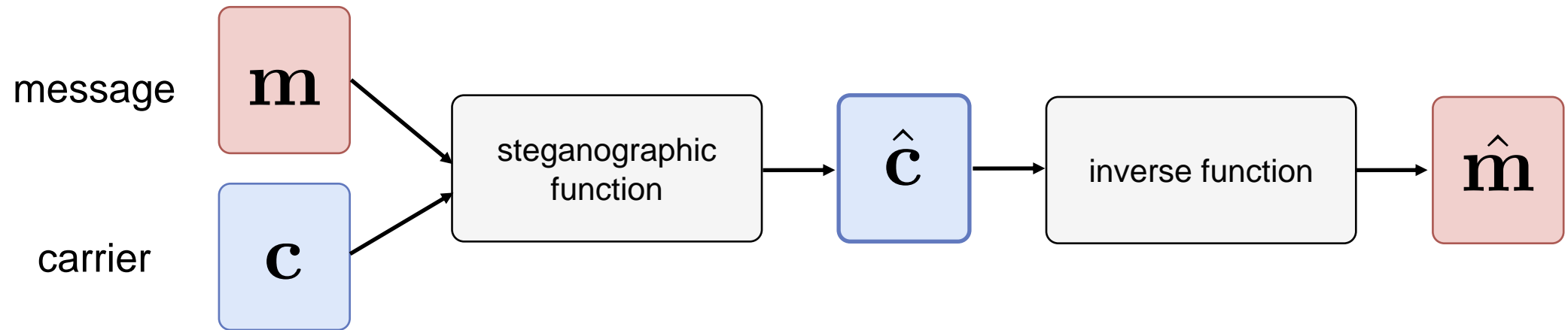
Aggregation by rate



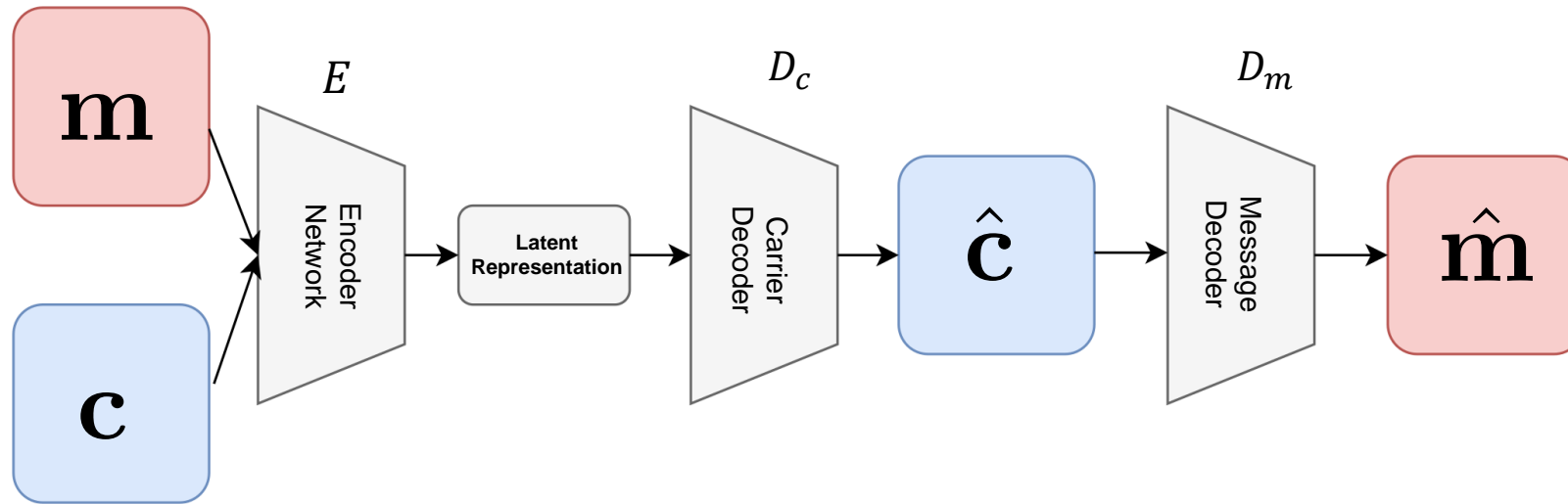
STEGAN - 0 - GRAPHY

(Secret Writing)

Steganography: problem setting

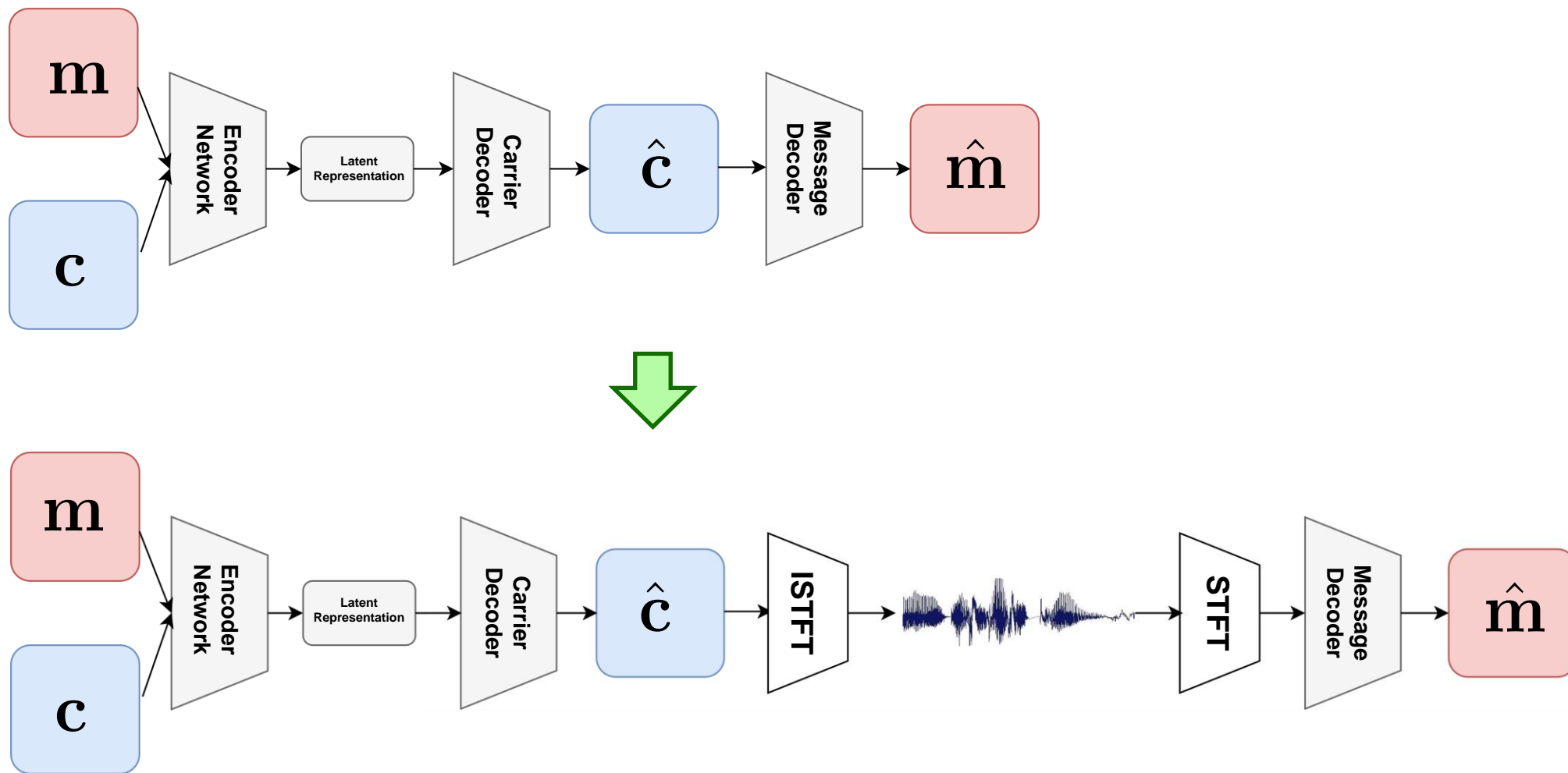


Steganography: model




$$\mathcal{L}(\mathbf{c}, \mathbf{m}) = \lambda_c \|\mathbf{c} - D_c(E(\mathbf{c}, \mathbf{m}))\|_2^2 + \lambda_m \|\mathbf{m} - D_m(D_c(E(\mathbf{c}, \mathbf{m})))\|_2^2$$

STFT+ISTFT within the network



Example 1

Original Carrier 

Reconstructed Carrier 


Original Message 


Reconstructed Message 

Example 2


Original Carrier 

Reconstructed Carrier 

Original Message 

Reconstructed Message 

Examples 3

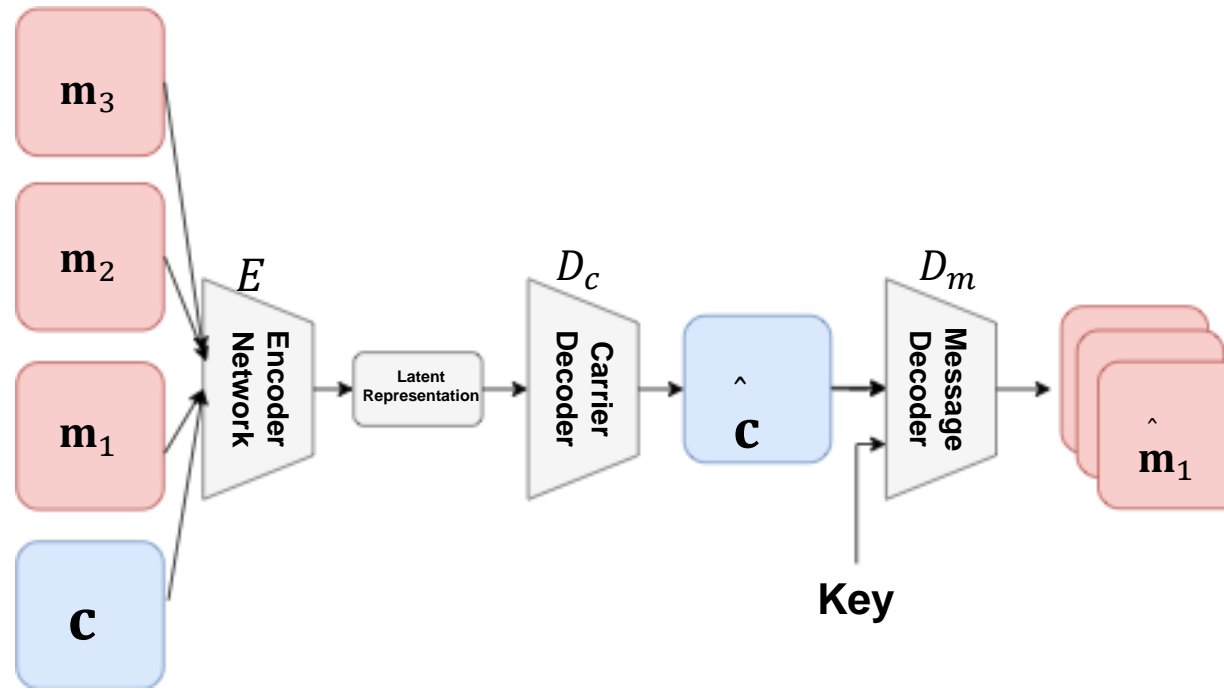
Original Carrier 

Reconstructed Carrier 

Original Message 

Reconstructed Message 

Conditional Decoder



$$\mathcal{L}(\mathbf{c}, \{\mathbf{m}_i\}_{i=1}^k) = \lambda_c \|\mathbf{c} - D_c(E(\mathbf{c}, \{\mathbf{m}_i\}_{i=1}^k))\|_2^2 + \lambda_m \sum_{i=1}^k \|\mathbf{m}_i - D_m(D_c(E(\mathbf{c}, \{\mathbf{m}_i\}_{i=1}^k)), q_i)\|_2^2$$

Thanks !!