Voice emotion transformation:

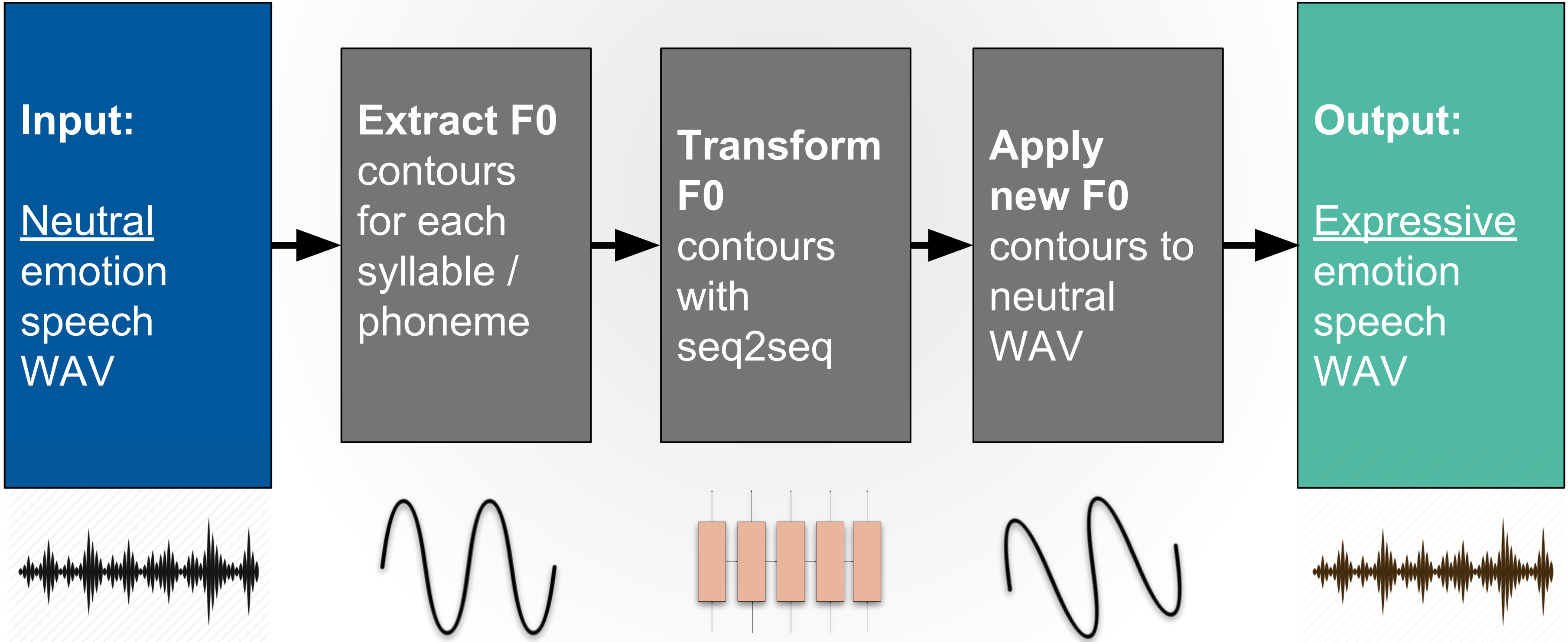
Generating expressive F0 contours with RNN-LSTM sequence-to-sequence models

Carl Robinson - TRIED M2 internship 2018 - IRCAM

Motivation

- F0 is the **vocal pitch** i.e. the *intonation* in the prosody
- Applications of F0 transformation
  - Voice assistants, screen readers...
  - Film, TV, video games
- My contribution:
  - Continuation of [Veaux & Rodet, 2011] (GMM-HMM)
  - Neural networks and sequence models

End-to-end Transformation Process



Input:

Neutral emotion speech WAV

Extract F0 contours for each syllable / phoneme

Transform F0 contours with seq2seq

Apply new F0 contours to neutral WAV

Output:

Expressive emotion speech WAV

Data

WAV Database:

Parallel database

Joy Anger Sad Fear

10 ph, 8 emo, 6 int = ~480 total

48KHz, 25ms, 5ms

Source/Target F0:

Neutral / Emotive

Phoneme contours

Variable-length

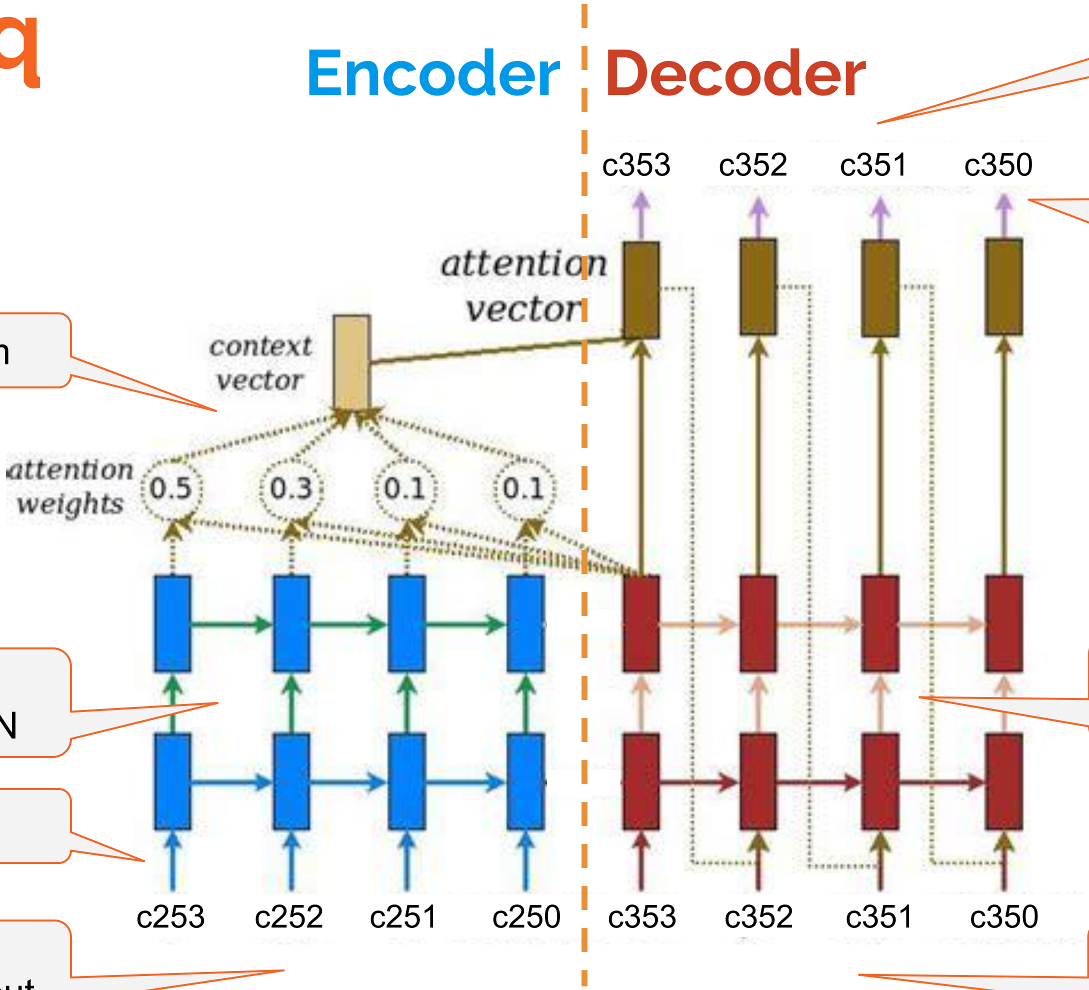
Integers

Listen:

Phrase 5:

- Original
- Joie (i02)
- Peur (i02)
- Tristesse (i02)
- Colère (i02)

Seq2seq



Output F0 contour

Softmax, ADAM, cross-entropy loss

Attention mechanism

Encoder: 128 cell, 2 layer bi-LSTM RNN

Decoder: 128 cell, 3 layer bi-LSTM RNN

Embedding layer

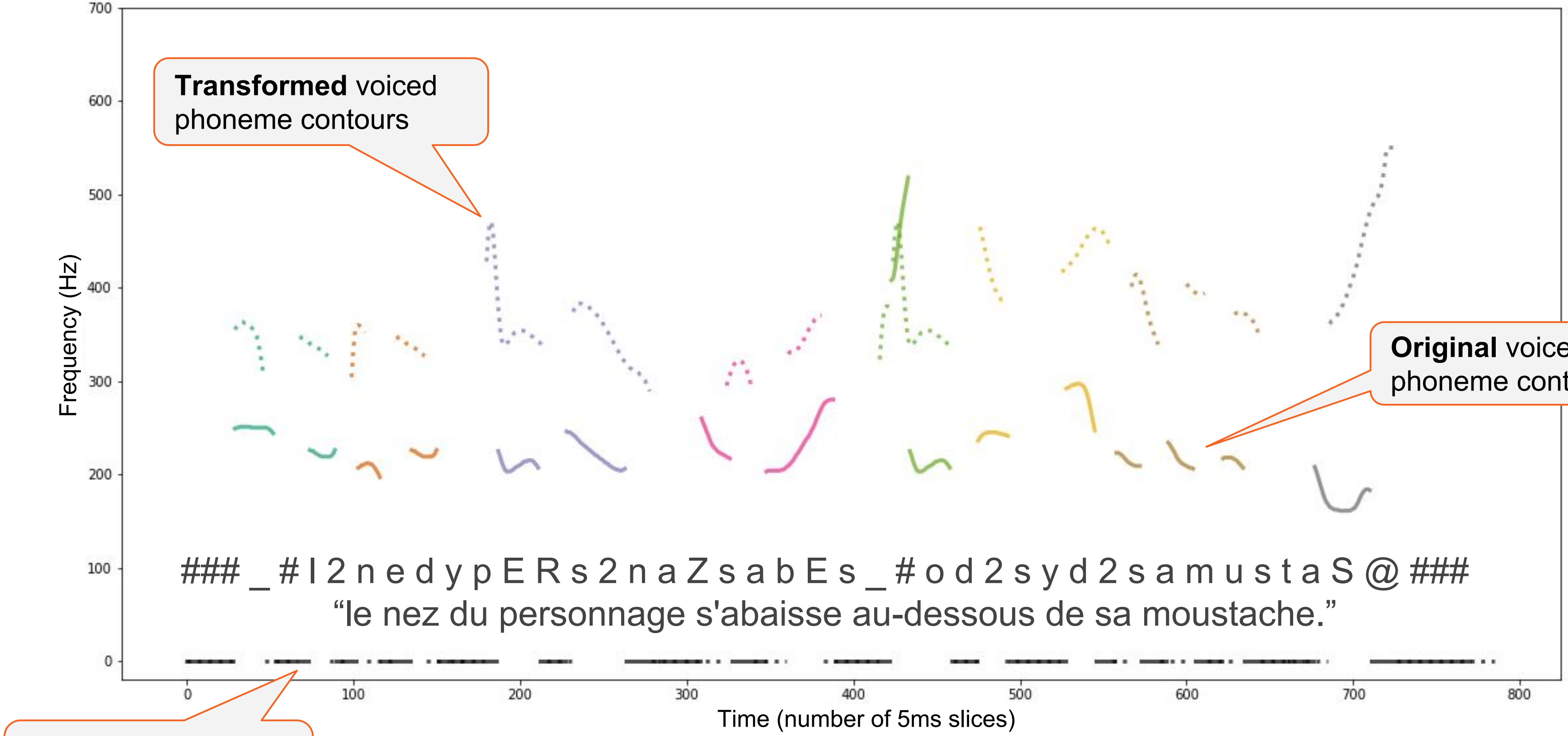
Syll-pos conditioned source phone F0 input

Syll-pos conditioned target phone F0 input

My contribution: technical details

- Phrase too long (4s/5ms = 800!) >> Split into **syllables**
- Syllables partly unvoiced >> Split into **voiced phonemes**
- Improve context >> **Match phrase & position**
- Improve context >> **Condition** F0 on syllable position

Phoneme frequency contours for a single phrase



Transformed voiced phoneme contours

Original voiced phoneme contours

Unvoiced phonemes and silences

Conditioned model

- Phrase 3 - Original
- Phrase 3 - Peur
- Phrase 3 - Colère

- Phrase 10 - Original
- Phrase 10 - Tristesse
- Phrase 10 - Peur

- Phrase 5 - Original
- Phrase 5 - Tristesse
- Phrase 5 - Joie

- Phrase 1 - Original
- Phrase 1 - Peur
- Phrase 1 - Joie

Non-conditioned model

- Phrase 5 - Original
- Phrase 5 - Joie
- Phrase 5 - Colère

- Phrase 10 - Original
- Phrase 10 - Peur
- Phrase 10 - Colère

- Phrase 3 - Original
- Phrase 3 - Tristesse
- Phrase 3 - Peur

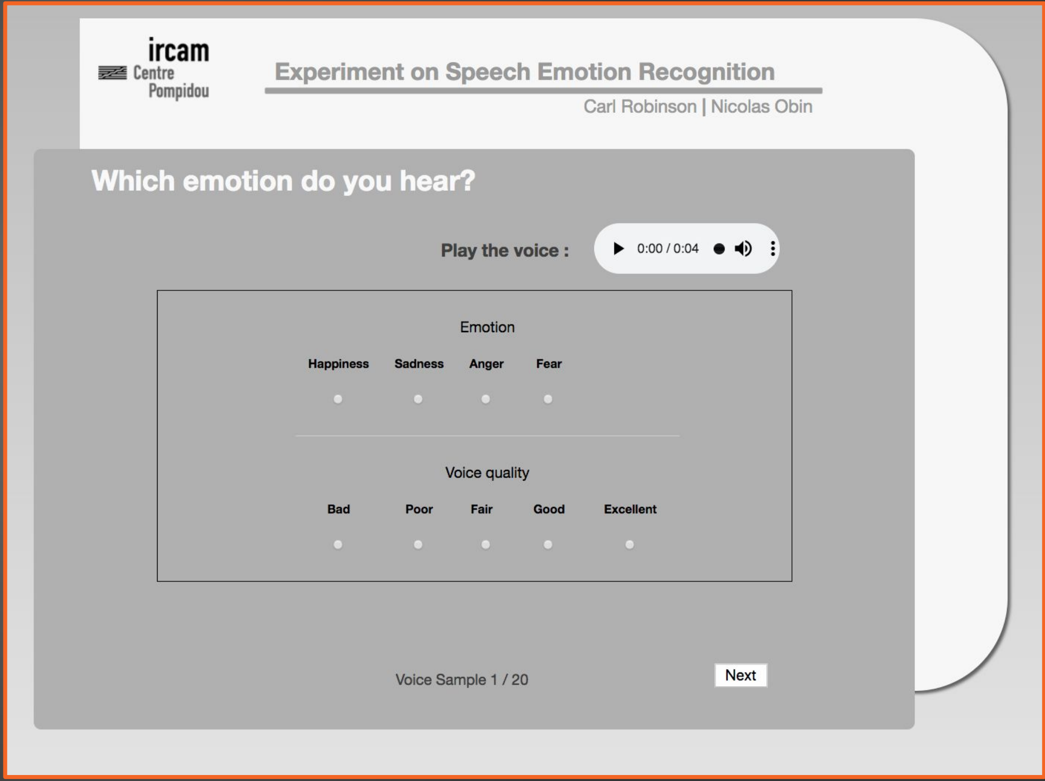
- Phrase 1 - Original
- Phrase 1 - Tristesse
- Phrase 1 - Joie

Experimental evaluation

- Joy / anger / sad / fear
- 96 samples:
  - 32 cond
  - 32 no-cond
  - 32 original
- 87 participants *

  20 randomised samples

  = 1734 responses



ircam Centre Pompidou

Experiment on Speech Emotion Recognition

Carl Robinson | Nicolas Obin

Which emotion do you hear?

Play the voice : ▶ 0:00 / 0:04

Emotion

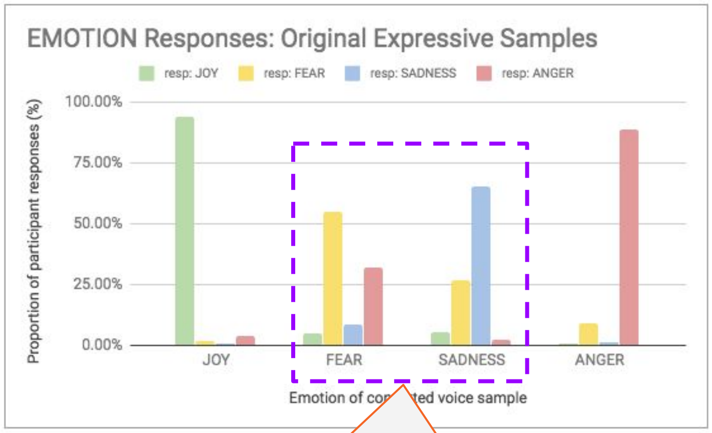
Happiness	Sadness	Anger	Fear
○	○	○	○

Voice quality

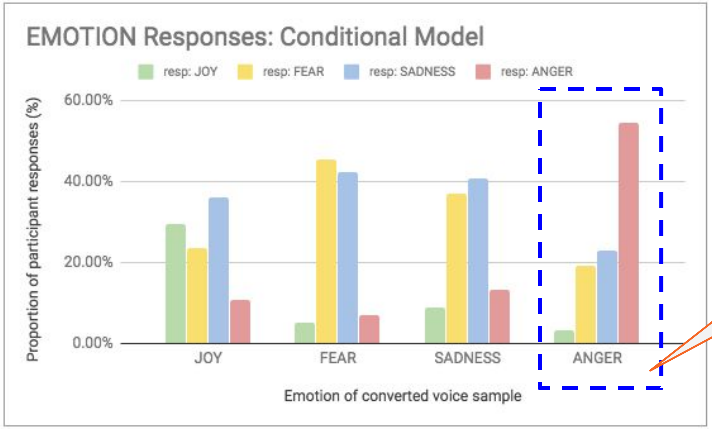
Bad	Poor	Fair	Good	Excellent
○	○	○	○	○

Voice Sample 1 / 20

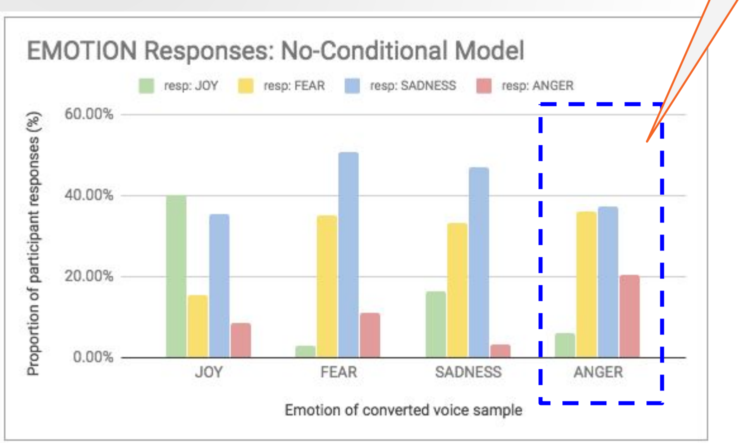
Results: Model Type



Originals samples are ambiguous; perception is mixed



Significant difference for anger between models

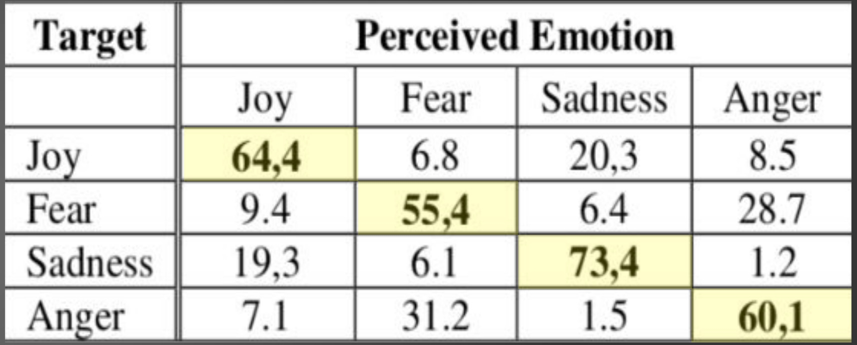


Comparison with Veaux & Rodet 2011

My best results

Target	Perceived Emotion				
	Joy	Fear	Sadness	Anger	Model
Joy	**40.27%**	15.44%	35.57%	8.72%	**NoCond**
Fear	5.00%	**45.63%**	42.50%	6.88%	**Cond**
Sadness	16.26%	33.33%	**47.15%**	3.25%	**NoCond**
Anger	3.21%	19.23%	23.08%	**54.49%**	**Cond**

Their best results



Target	Perceived Emotion			
	Joy	Fear	Sadness	Anger
Joy	64,4	6.8	20,3	8.5
Fear	9.4	55,4	6.4	28.7
Sadness	19,3	6.1	73,4	1.2
Anger	7.1	31.2	1.5	60,1

- Same parallel dataset used
- Best model for Fear & Anger = Conditioned on syll position
- Best model for Joy & Sadness = Not Conditioned
- My output sequence lengths generated, not forced !

Conclusions

- Seq2seq transforms F0 intonation as well as prev. work
- More context = better results (same phrase, same phon)
- Conditioning on syll pos only benefits some emotions

Future work

- Larger dataset, multiple speakers & genders
- Model other vocal components in addition to frequency
- Multi-tier architecture for syll/phrase level correlations

Thank you

carl.robinson@gmail.com

ircam
Centre
Pompidou

TELECOM
SudParis

le cnam

UNIVERSITÉ DE
VERSAILLES
ST-QUENTIN-EN-YVELINES
université PARIS-SACLAY