Hankook University of Foreign Studies
ELLT Graduate School Colloquium
May. 12, 2022

Modeling Emotional States from Speech Expressions

Tae-Jin Yoon
Department of English Language and Literature
성신여자대학교 SUNGSHIN UNIVERSITY

CONTENTS

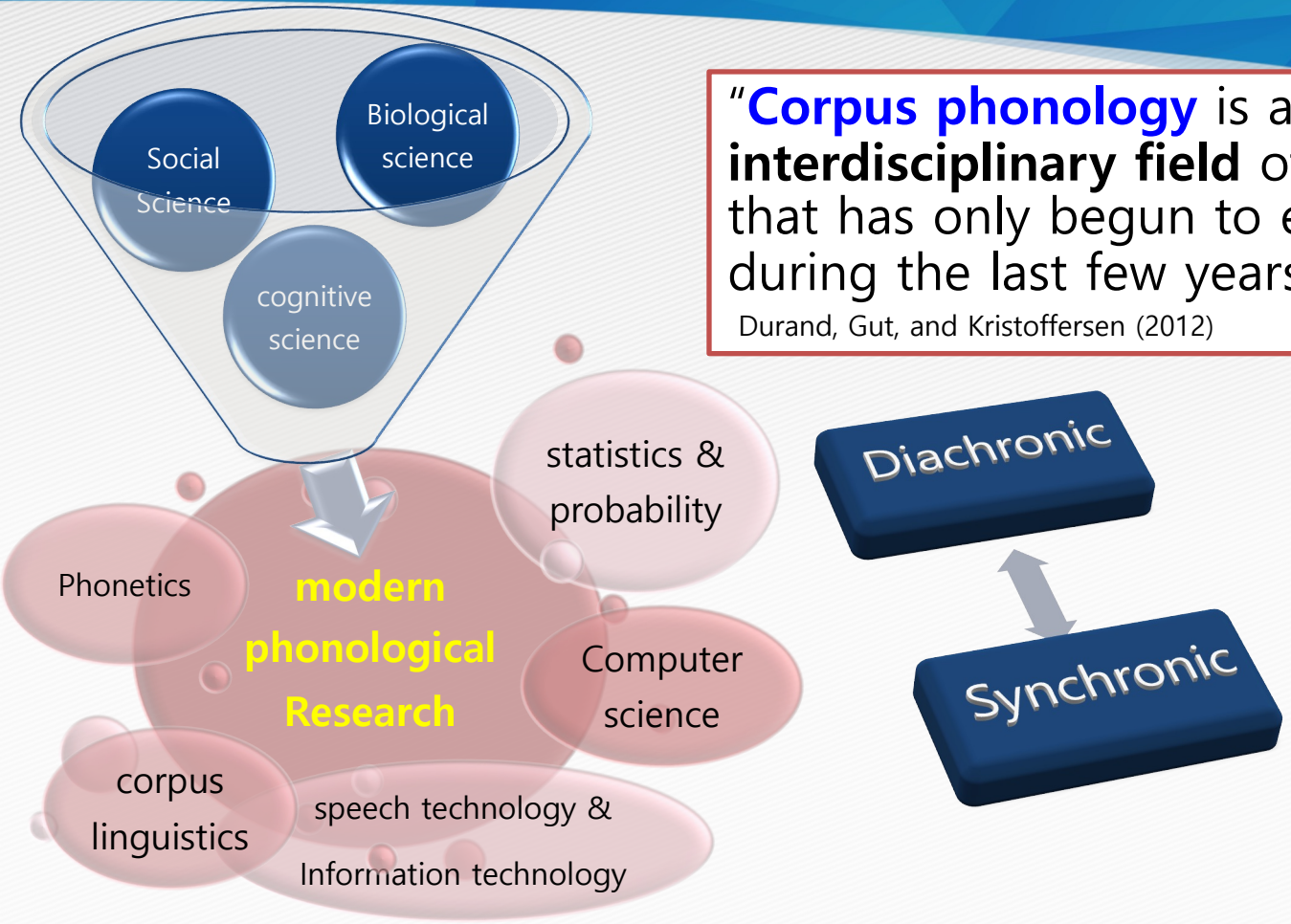
Challenges of modern phonological research

- Without the ability to collect and analyze speech sounds in various environments and contexts, it can be a significant obstacle to trying to do the latest phonetics/phonology research.
- Learning a research methodology that can collect and process a variety of phonological and meta-information and speech data would be necessary to participate in various academic discussions, which is not an easy task to learn in a short time.

-카이스트가 왜 인문학 위기를 걱정하나.

"우리 사회엔 분명 인문학이 필요하다. 앞으론 문·이과 융합 인재 수요가 폭발적으로 늘어날 거다. 스스로 새 사상을 창조하는 사람이 리더가 된다. 대표적인 분야가 'AI와 협력'이다. 지난 2017년 바둑기사가 AI와 한 조를 이뤄 대결하는 복식경기에서 중국 롄샤오 8단이 구리 9단을 상대로 예상 밖 불계승을 거뒀다. 일대일로 붙으면 8단 기사가 지겠지만, AI와 협력을 더 잘한 롄샤오가 구리를 이겼다. 바둑뿐 아니라 앞으로 모든 영역에서 AI와 협력을 잘하는 방법을 연구하는 인문학이 필요할 것이다. 인문학도 4차 산업혁명 시대엔 연구 툴(방식)을 바꿔야 한다. 지금의 인문계 교육은 '문송(문과라서 죄송)'은 사라지지 않는다."

https://www.chosun.com/national/weekend/2022/04/30/LBHZPEH7YRFTFIC3KU7GFEJVLU/?utm_source=naver&utm_medium=referral&utm_campaign=naver-news



"**Corpus phonology** is a new **interdisciplinary field** of research that has only begun to emerge during the last few years."

Durand, Gut, and Kristoffersen (2012)

In *The Handbook of Corpus Phonology* (2012)

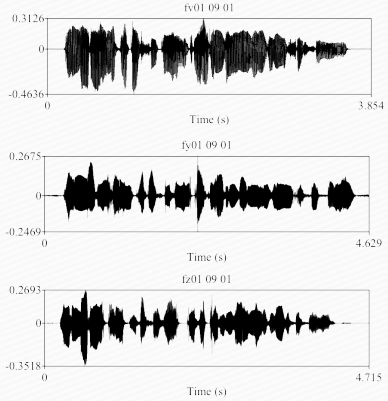
Opportunities

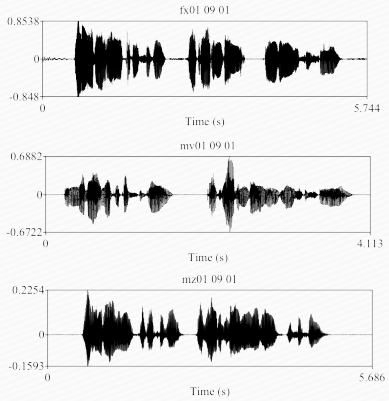
- Several benefits to doing phonetic research on a large database
  - the ability to do robust testing of linguistic hypotheses
  - the use of data from naturalistic interactions
  - the possibility of emergence of macroscopic structure in large enough databases.

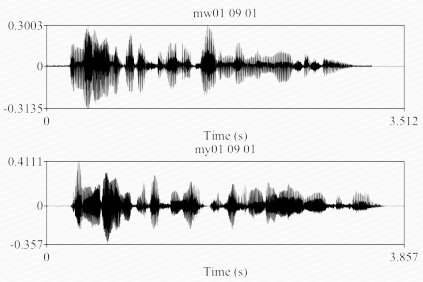
(Goldstein, 2011)

Challenges

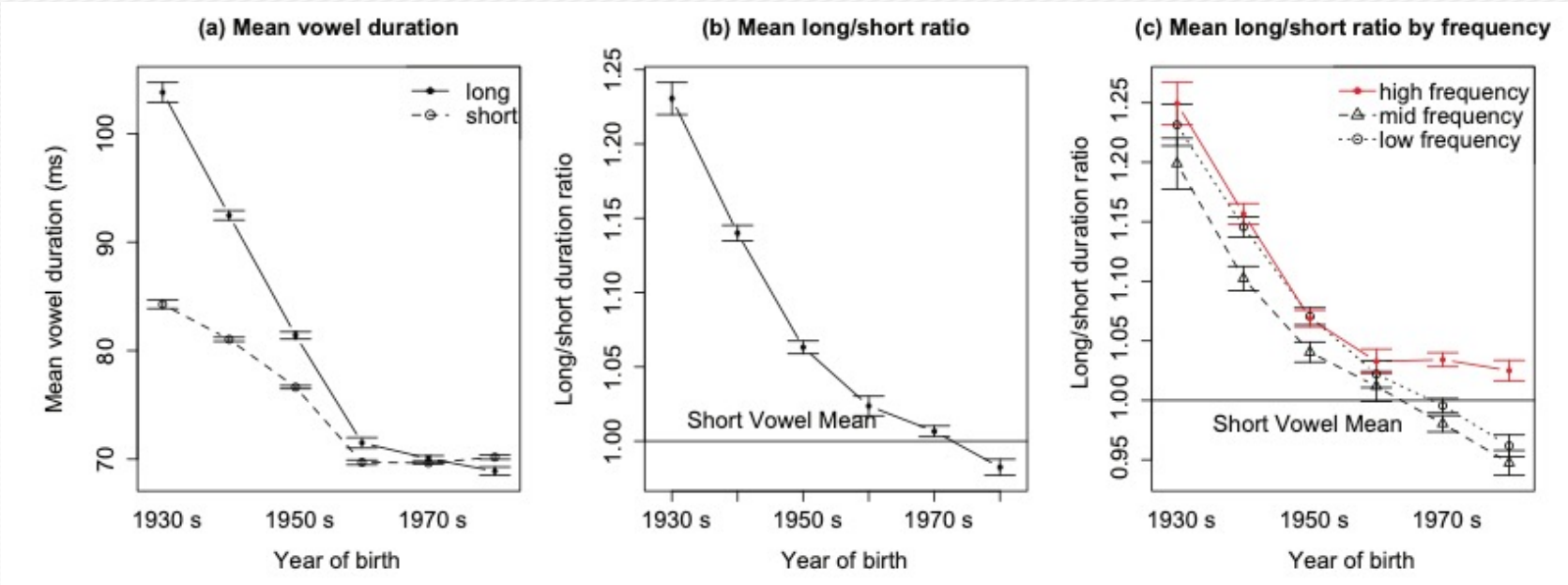
- How can we (semi)automatically extract the kinds of representation and features from large corpora?
- What kinds of tests can we apply to the extracted features to test our phonologically motivated questions?





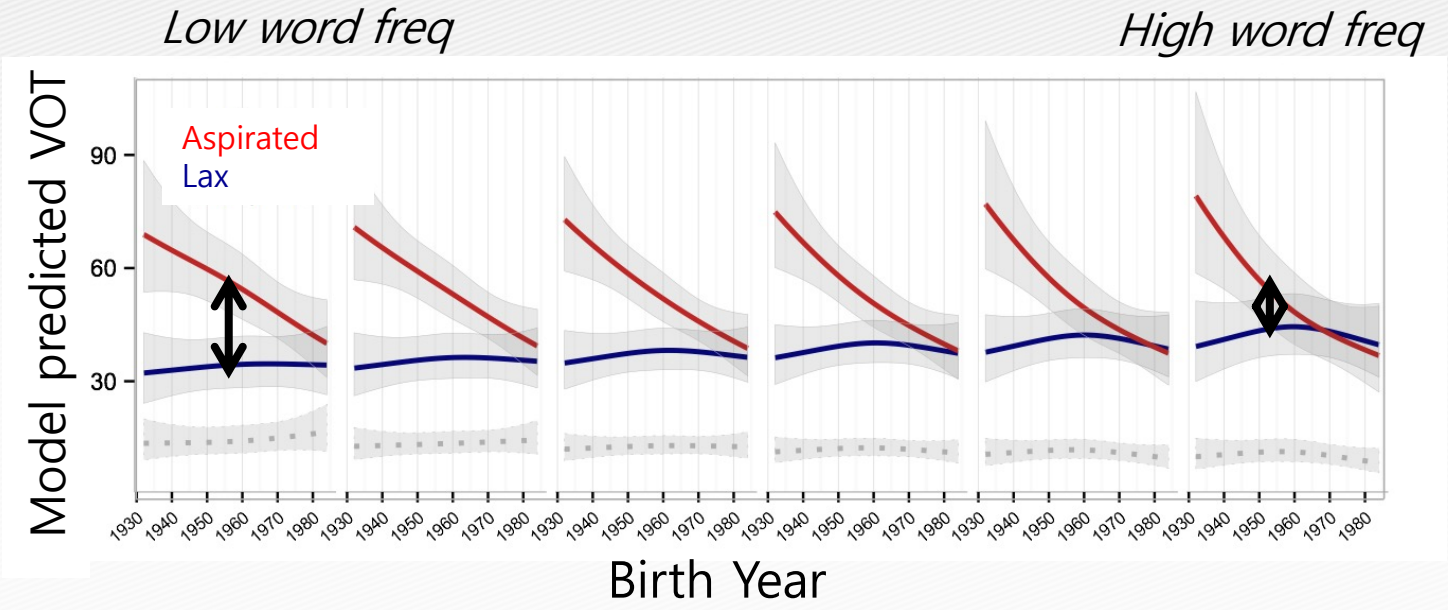


Vowel Length merger in Seoul Korean



Kang, Yoonjung, Tae-Jin Yoon, and Sungwoo Han (2015). Frequency effects on the vowel length merger in Seoul Korean. *Laboratory Phonology*, 6(3-4): 469–503.

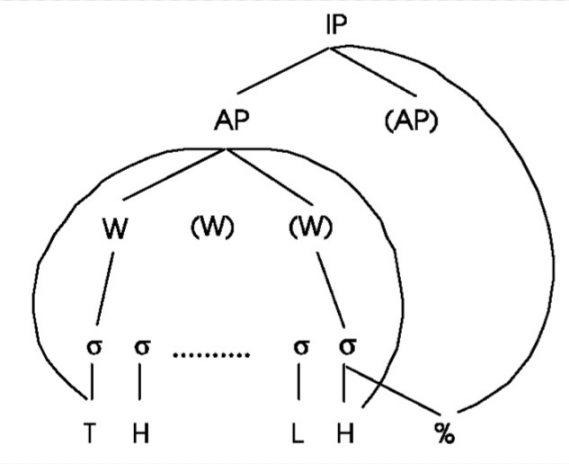
VOT and frequency over time



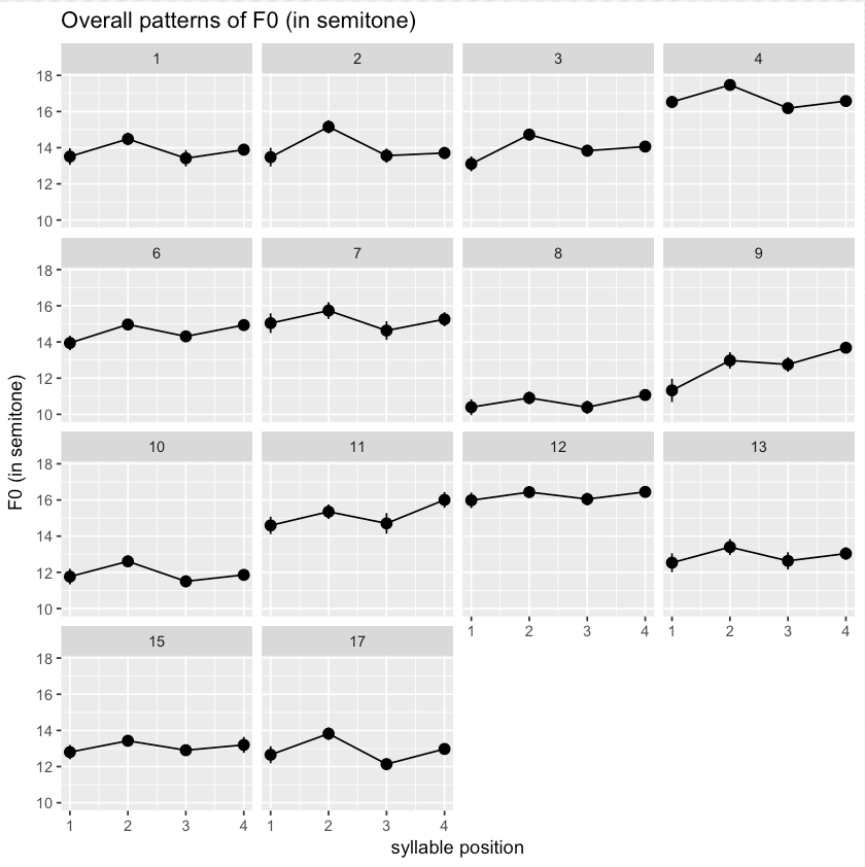
Higher frequency words: smaller VOT difference (p =0.025)

Hye-Young Bang, Morgan Sonderegger, Yoonjung Kang, Megan Clayards, and Tae-Jin Yoon (2018) The emergence, progress, and impact of sound change in progress in Seoul Korean: implications for mechanisms of tonogenesis. *Journal of Phonetics*. 66. 120-144.

Modeling of Nucleus F0 on Korean Accentual Phrase

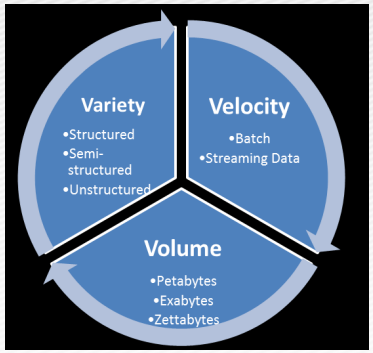


Jun, S.-A. (2005)



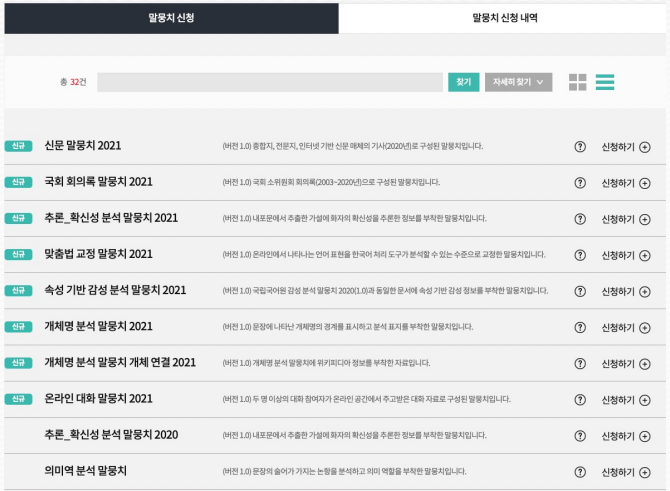
Tae-Jin Yoon (2017) Growth Curve Modeling of Nucleus F0 on Korean Accentual Phrase. *Phonetics and Speech Sciences* 9(3), 17-23.

Corpus



LDC Linguistic Data Consortium

AI Hub



https://www.researchgate.net/figure/BM-big-data-characteristics-3V-Adopted-from-4_fig6_303562879

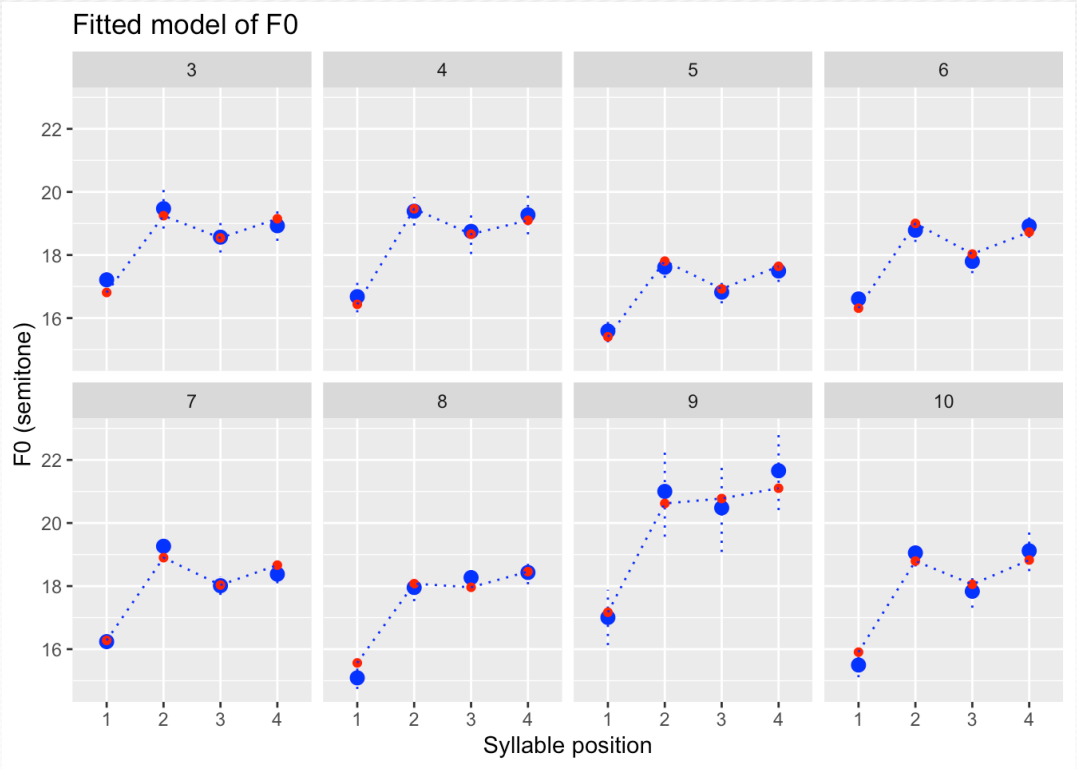
Children's developmental pattern of AP

AI Hub

음성/자연어

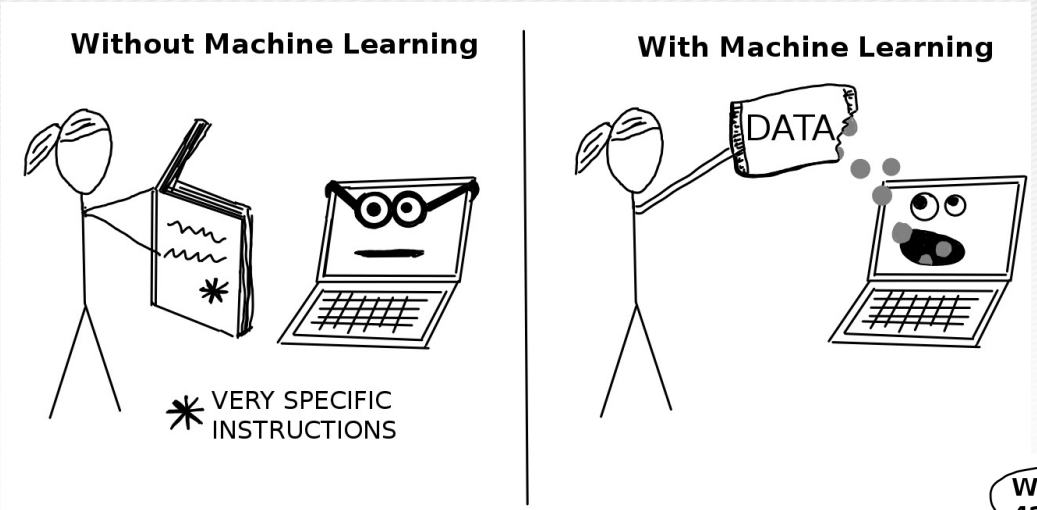
자유대화 음성(소아, 유아)

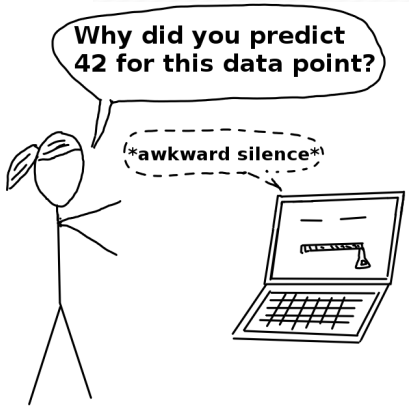
텍스트 오디오 2020



Modeling Emotional States from Speech Expressions

Data-driven approach





https://christophm.github.io/interpretable-ml-book/terminology.html

Why Emotion in phonetics?

- Emotions: psychological states variously associated with thoughts, feelings, behavioral responses, and a degree of pleasure or displeasure.

Source: https://en.wikipedia.org/wiki/Emotion

- Modulation of pitch, loudness, duration, and voice quality across syllables in an utterance

→ conveys both linguistic and non-linguistic information

prominence, prosodic phrasing, …

age, gender, speaker's emotional status, etc…

Two theories on emotion

- **The discrete emotion theory**
  - Basic discrete emotions

    (1) surprise, (2) interest, (3) joy, (4) rage, (5) fear, (6) disgust, (7) shame, (8) anguish
  - Individual emotions have biological and neurological profiles
- **The dimensional theory**
  - Two emotional dimensional spaces distinguish emotions

    (1) **valence** – how positive or negative an emotion is

    (2) **arousal** – the intensity of an emotion

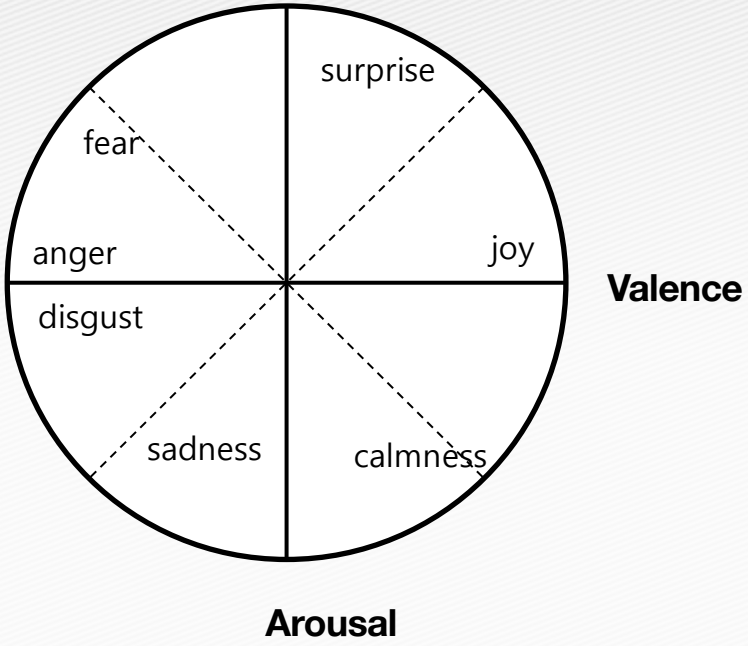
The discrete emotion approach

- ❑ Emotions are discrete, measurable, and physiologically distinct.
- ❑ Certain emotions appeared to be universally recognized.

→ Many studies have examined the vocal characteristics of speech in hope of defining a vocal signature for each basic emotion (Russell 2003)

The Dimensional approach

- So far, the strongest single association found for vocal acoustic have been with the sender's general **arousal** level.
- High-arousal emotions such as **anger** and **joy** have similar characteristics low arousal emotions such as **sadness**
  - greater loudness,
  - higher pitch
  - faster speech
- Few works have concentrated on distinguishing emotions between positive- and negative- valence emotions such as **anger** and **joy**.



Eerola, T., & Vuoskoski, J. K. (2010). *A comparison of the discrete and dimensional models of emotion in music. Psychology of Music, 39(1), 18–49.* doi:10.1177/0305735610362821

Research topic

- F0 contours (not a summary F0 such as meanF0 or F0 range) contains discriminatory information about emotions.
- Very few can be found in the literature that made the efforts to describe the shape of f0 contours directly in classifying emotions

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

- The RAVDESS dataset is a multimodal validated English dataset that contains speech, song, and video files that represent 8 emotions.
- The portion of the dataset that I use in this study is the speech audio files that are represented by 1440 wave file.
- Twenty-four professional actors (12 female and 12 male) with 60 trials for each actor produced the 1440 wave files (24×60 =1440).

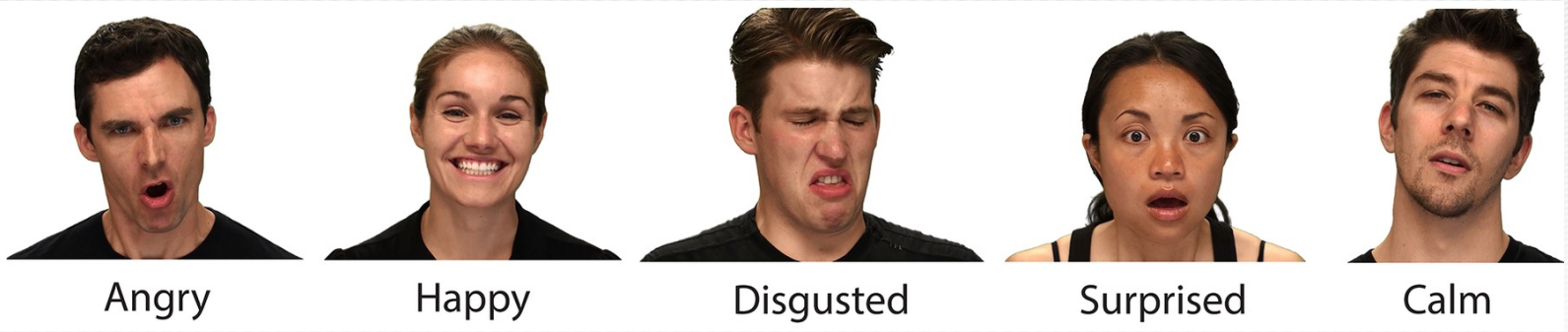
	Fname	Emotions	Intensity	Repetition	Actor	Gender	Statement
0	03-01-05-01-02-01-16	angry	normal	1	16	female	Dogs are sitting by the door
1	03-01-06-01-02-02-16	fear	normal	2	16	female	Dogs are sitting by the door
2	03-01-06-02-01-02-16	fear	strong	2	16	female	Kids are talking by the door
3	03-01-05-02-01-01-16	angry	strong	1	16	female	Kids are talking by the door
4	03-01-07-01-01-01-16	disgust	normal	1	16	female	Kids are talking by the door

Livingstone, S.R. and F.A. Russo. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE* 13(5), e0196391.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (2)

- The actors vocalized two sentences in a neutral North American accent.
  - “Kids are talking by the door”
  - “Dogs are sitting by the door”
- The emotions included in this dataset are
  - neutral, calm, happy, sad, angry, fearful, surprise, and disgust
- Each expression is produced at two levels of emotional intensity (normal and strong) except for the neutral emotion that is recorded in a normal intensity only.

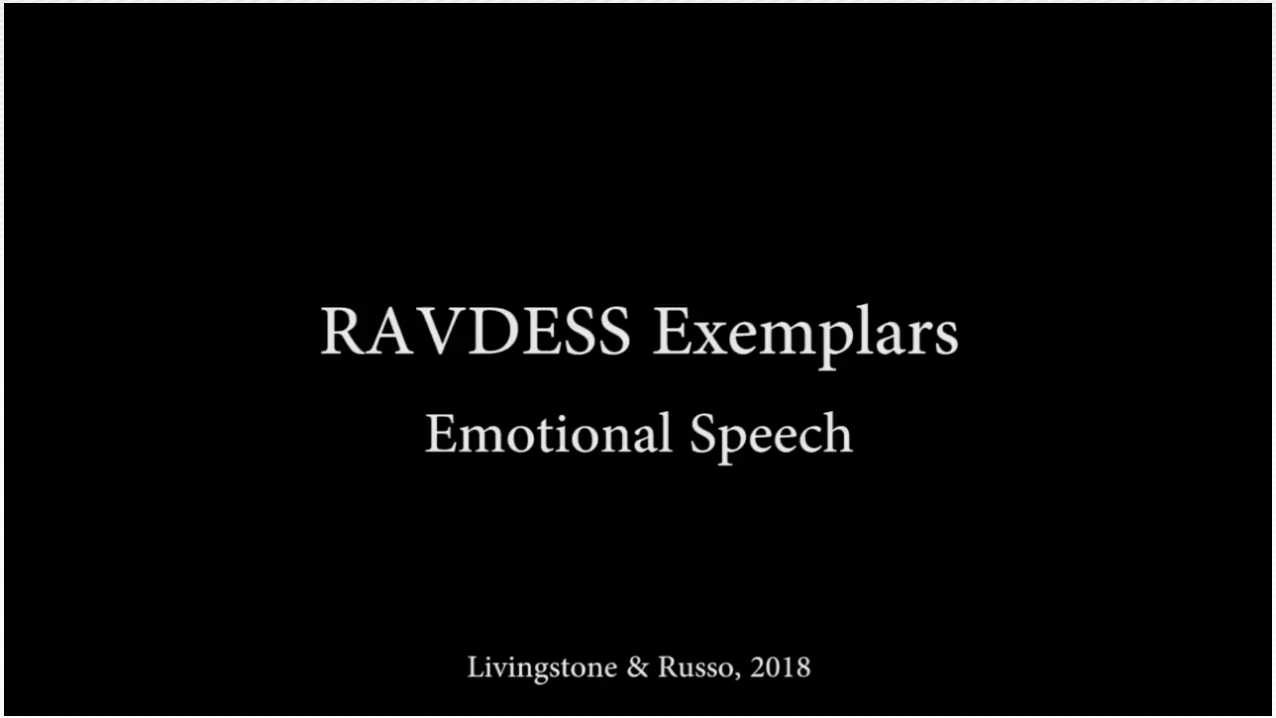
The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (3)



neutral calm happy sad angry fearful disgust surprised

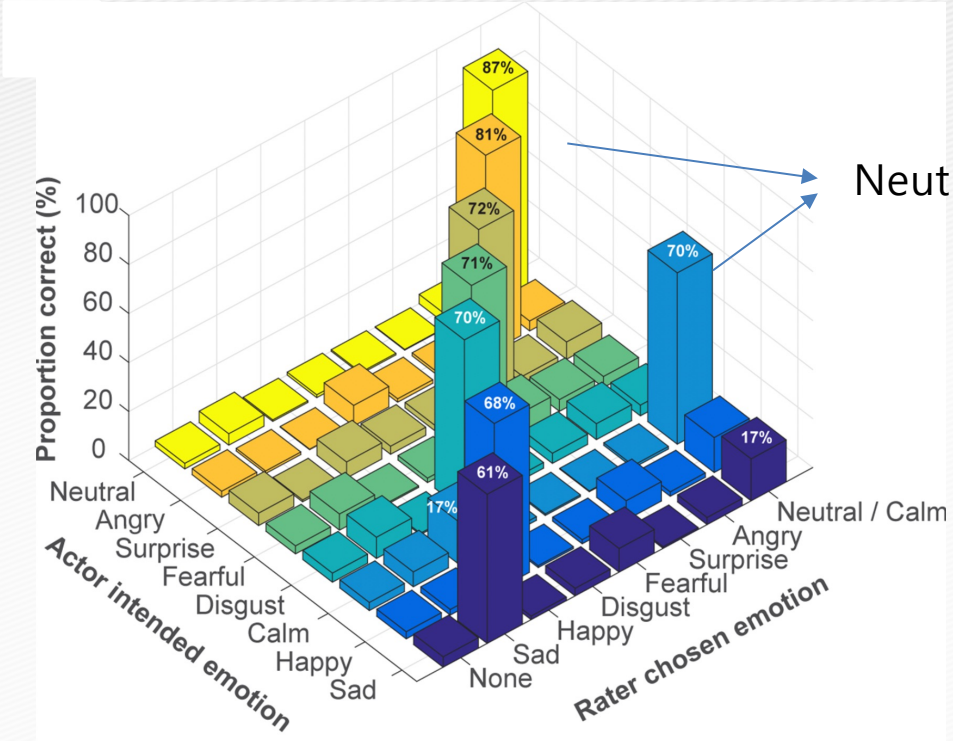
https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

RAVDESS Examplars



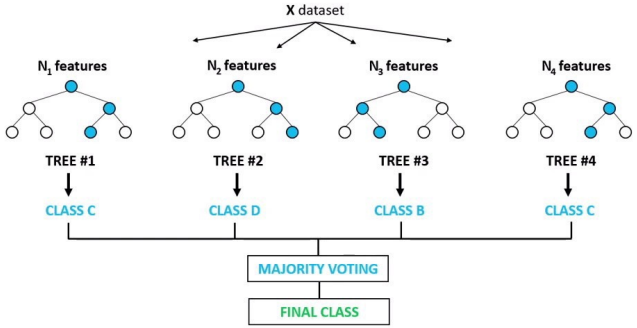
Validity of the RAVDESS dataset

- 247 untrained research participants from North America



Livingstone, S.R. and F.A. Russo. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE* 13(5), e0196391.

I. Random Forest



	precision	recall	f1-score	support
neutral	0.40	0.16	0.23	25
calm	0.47	0.79	0.59	61
happy	0.49	0.30	0.37	66
sad	0.45	0.30	0.36	66
angry	0.72	0.60	0.65	72
fearful	0.53	0.48	0.50	60
disgust	0.37	0.52	0.43	60
surprised	0.49	0.59	0.54	66
accuracy			0.49	476
macro avg	0.49	0.47	0.46	476
weighted avg	0.50	0.49	0.48	476

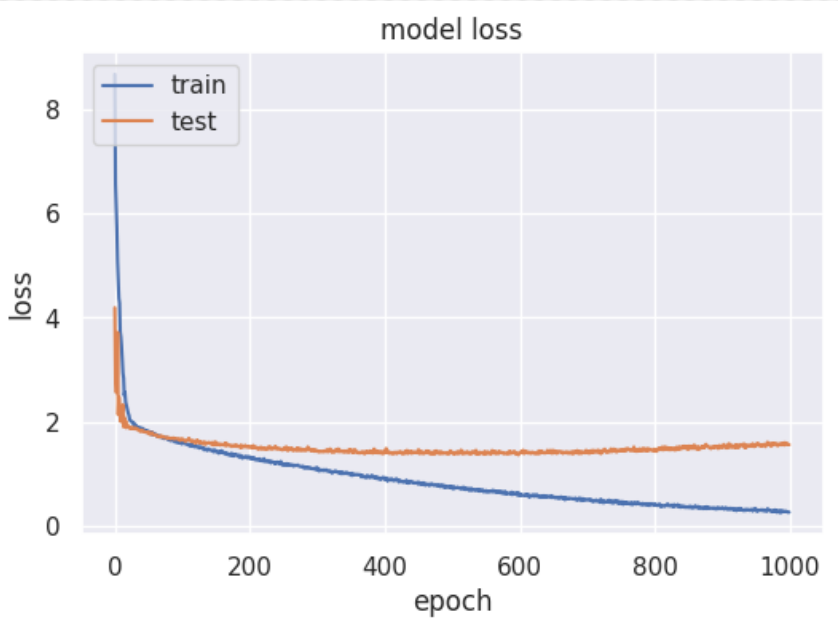
Support = # of samples

- Precision: 양성으로 예측된 것 중 진짜 양성은 얼마나?
- Recall: 양성 샘플 중 양성 클래스로 분류된 샘플의 수는?
- F-score: precision과 recall의 조화 점수

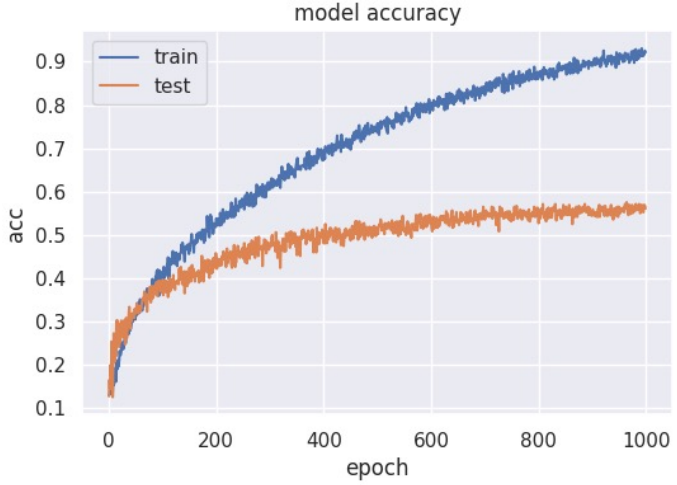
Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.".

II. Deep Learning (CNN)

```
Model: "sequential_13"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_56 (Conv1D)           (None, 40, 128)           768
_________________________________________________________________
activation_69 (Activation)   (None, 40, 128)           0
_________________________________________________________________
dropout_26 (Dropout)         (None, 40, 128)           0
_________________________________________________________________
max_pooling1d_18 (MaxPooling (None, 5, 128)            0
_________________________________________________________________
conv1d_57 (Conv1D)           (None, 5, 128)            82048
_________________________________________________________________
activation_70 (Activation)   (None, 5, 128)            0
_________________________________________________________________
dropout_27 (Dropout)         (None, 5, 128)            0
_________________________________________________________________
flatten_13 (Flatten)         (None, 640)               0
_________________________________________________________________
dense_13 (Dense)             (None, 8)                 5128
_________________________________________________________________
activation_71 (Activation)   (None, 8)                 0
=================================================================
Total params: 87,944
Trainable params: 87,944
Non-trainable params: 0
```



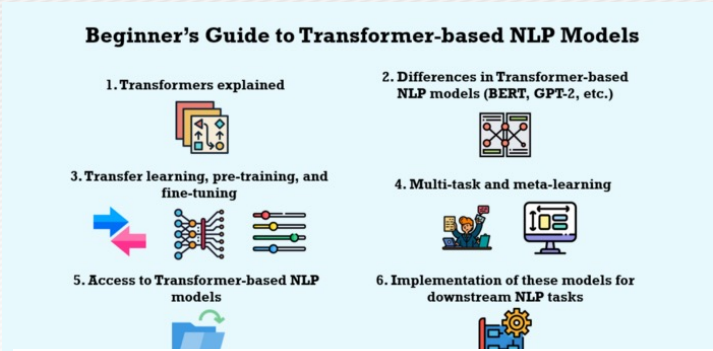
modified from https://github.com/marcogdepinto/emotion-classification-from-audio-files/blob/master/legacy_code/data_exploration/EmotionsRecognition.ipynb

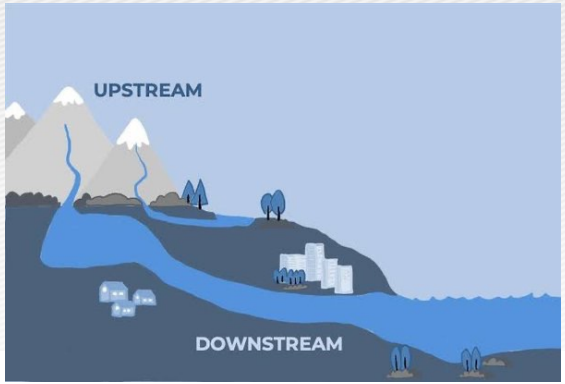


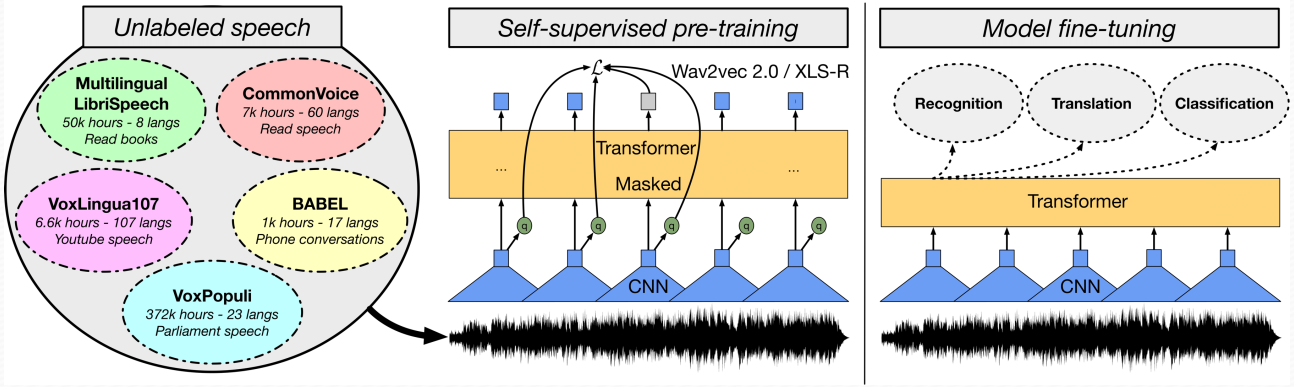
	precision	recall	f1-score	support
neutral	0.26	0.40	0.31	25
calm	0.65	0.59	0.62	61
happy	0.53	0.61	0.56	66
sad	0.58	0.45	0.51	66
angry	0.70	0.60	0.65	72
fearful	0.58	0.63	0.61	60
disgust	0.50	0.55	0.52	60
surprised	0.60	0.56	0.58	66
accuracy			0.56	476
macro avg	0.55	0.55	0.55	476
weighted avg	0.58	0.56	0.57	476

de Pinto, M. G., Polignano, M., Lops, P., & Semeraro, G. (2020, May). Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients. In *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)* (pp. 1-5). IEEE.

III. Transfer Learning

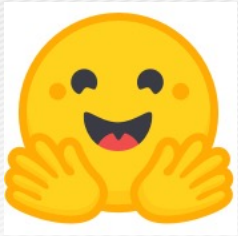






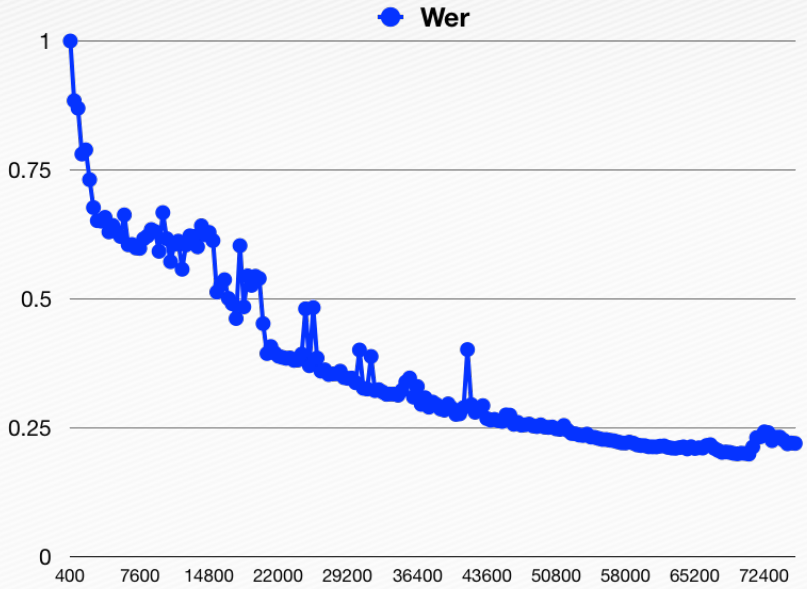
Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449-12460.

wav2vec2.0 for Children-Specific ASR



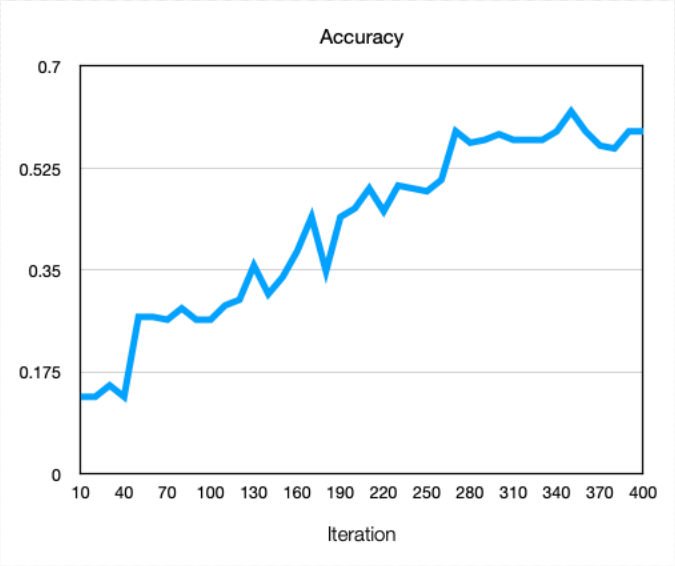
```
from unicode_jamo import *
join_jamos(processor.decode(pred_ids))
```

'거북인 이렇게 겁자난 엉그멍금'




```
***** train metrics *****
  epoch                    =                 30.0
  total_flos               =       465552446523GF
  train_loss               =               0.4268
  train_runtime            = 14 days, 11:02:18.34
  train_samples            =               161732
  train_samples_per_second =                3.884
  train_steps_per_second   =                0.061
```

wav2vec2.0 based classification

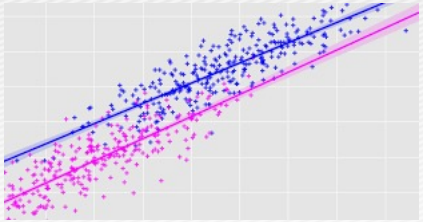


	precision	recall	f1-score	support
angry	0.73	0.81	0.77	27
calm	0.84	0.78	0.81	27
disgust	0.72	0.67	0.69	27
fear	0.37	0.82	0.51	28
happy	0.50	0.07	0.13	27
neutral	0.00	0.00	0.00	14
sad	0.26	0.33	0.30	27
surprise	0.83	0.74	0.78	27
accuracy			0.56	204
macro avg	0.53	0.53	0.50	204
weighted avg	0.57	0.56	0.53	204

Support = # of samples

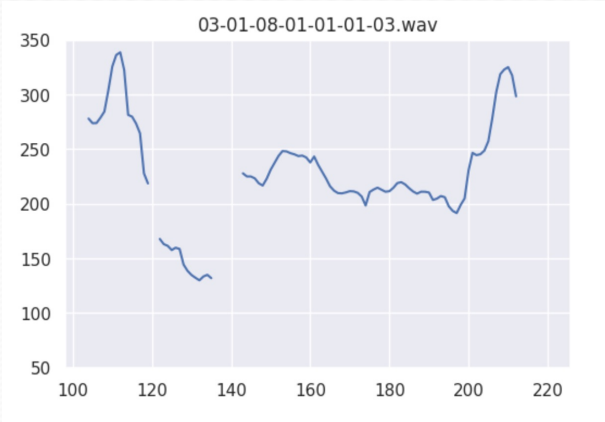
IV. Generalized Additive Mixed Modeling

- In Linear Model, the mean of data is modeled as a sum of linear terms



$$y_i = \beta_0 + \sum_j \beta x_{ji} + \varepsilon_i$$

- In Generalized Additive Mixed Model, the mean of data is modeled as a sum of ***smooth*** functions (= smooths)

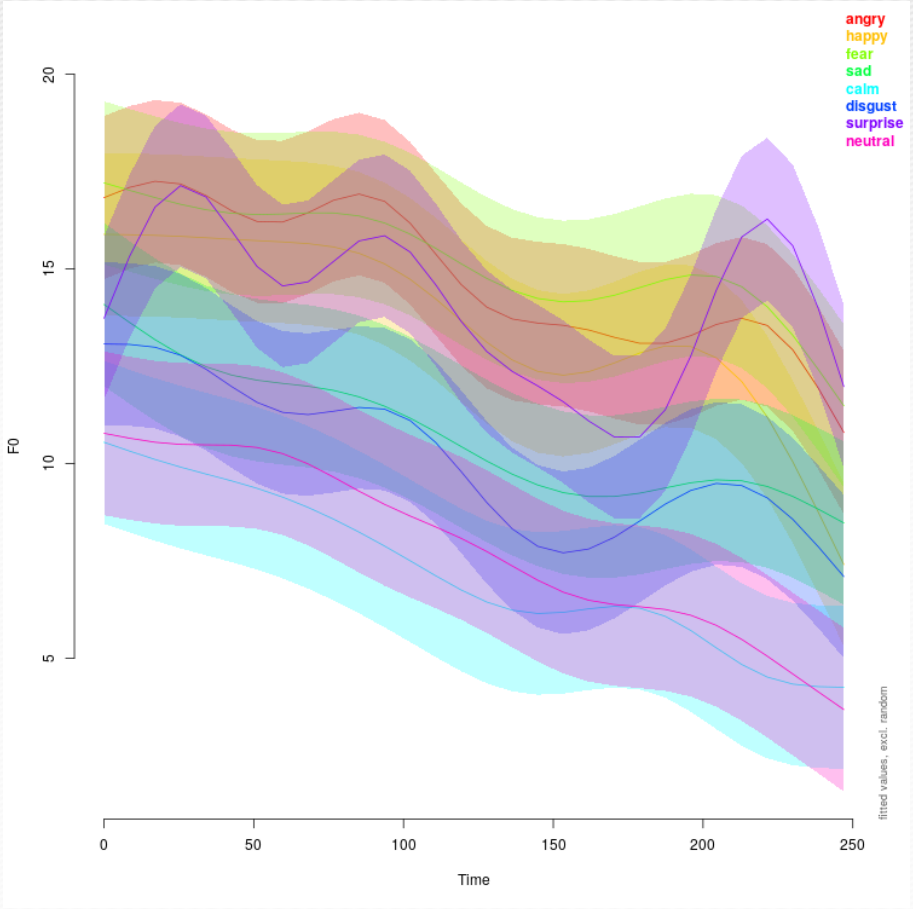


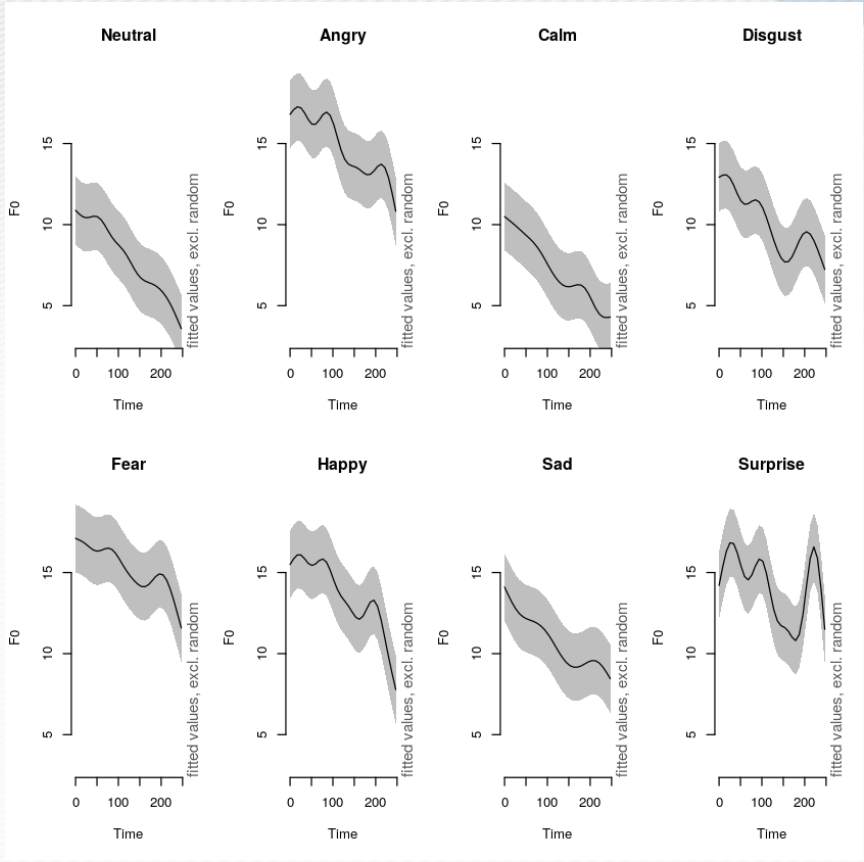
$$y_i = \beta_0 + \sum_j s_j(x_{ji}) + \varepsilon_i$$

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction. *arXiv preprint arXiv:1703.05339*.

GAMM approach to the F0 contour modeling





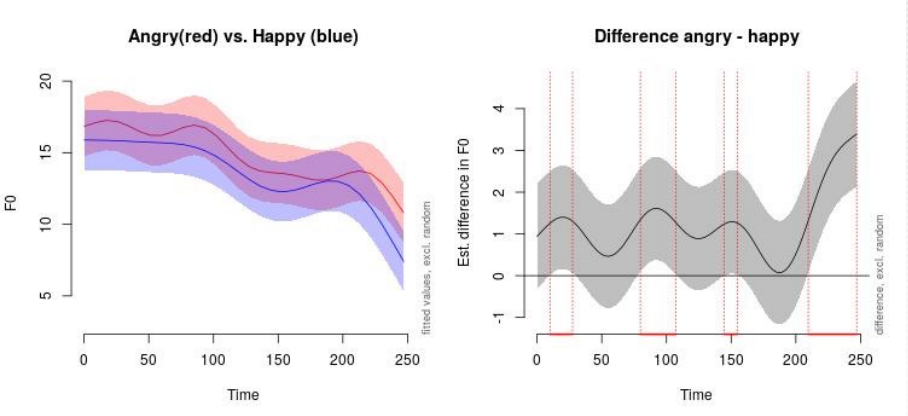
Gamm Modeling

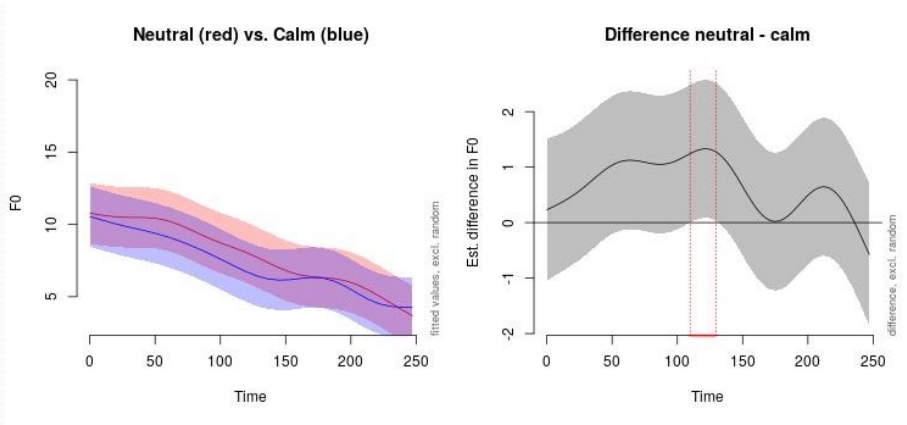
```
Formula:
F0 ~ Emotions + s(Time, by = Emotions, k = 10) + s(Actor, bs =
"re") +
    s(Actor, Emotions, bs = "re")

Parametric coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.9017     1.0637   7.429 1.10e-13 ***
Emotionsangry      6.9618     0.6277  11.091  < 2e-16 ***
Emotionscalm      -0.6860     0.6277  -1.093 0.274459
Emotionsdisgust    2.2108     0.6277   3.522 0.000428 ***
Emotionsfear       7.3391     0.6277  11.692  < 2e-16 ***
Emotionshappy      5.7515     0.6277   9.163  < 2e-16 ***
Emotionssad        2.8123     0.6277   4.480 7.46e-06 ***
Emotionssurprise   6.2753     0.6277   9.997  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.649   Deviance explained = 64.9%
fREML = 1.0445e+06  Scale est. = 20.228    n = 357120
```

Pair-wise comparison of contours





- F0 contour differentiates angry from happy.
- F0 contour hardly differentiates calm from neutral.

Conclusion

- I reviewed briefly what and how we can do with large-scaled datasets of spoken language.
- I attempted to compare approaches to emotional classification
- I put an emphasis on modelling emotions using F0 contours as an input to generalized additive model (GAM)
  - The present approach has predictive power.
  - The additive model provides visualized aids and makes us better understand validity of the data obtained from human labelers.

Thank you