CS578: DIGITAL SPEECH SIGNAL PROCESSING

SPEECH EMOTION RECOGNITION

MODELING, FEATURES, AND CLASSIFIERS

Dr George Kafentzis
Lecturer
Computer Science Department
University of Crete

OVERVIEW

- **Introduction**
- Computational Modeling of Emotion
- Acoustic Features
- Classifiers
- Conclusions

INTRODUCTION

- Relatively recent research field: speech emotion recognition
  - Defined as extracting the emotional state of a speaker from his or her speech.

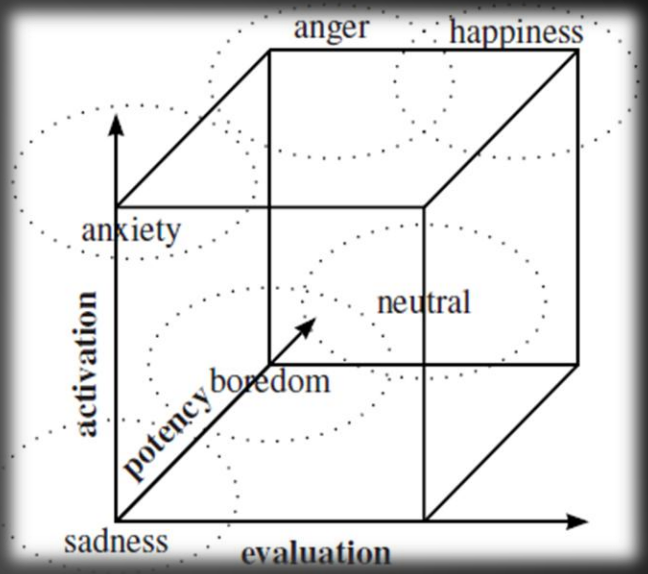
- Useful for applications which require natural man–machine interaction
  - web movies and computer tutorial applications [1]
  - in-car board system [1]
  - diagnostic tools for therapists [2]
  - aircraft cockpits [3]
  - call center applications [4]
  - mobile communication [4]
  - etc.

INTRODUCTION

- Abundance of studies on emotion/affect and speech
  - First studies in the early decades of the 20th century [5, 6]
  - First experiments on durational and pitch characteristics of simulated emotions [7, 8]
- In the 1990s, automatic speech processing began to deal with emotions
  - Pattern recognition techniques using prosodic features to classify the emotional content of utterances [9]
    - Reported performance close to human performance!
- At the turn of the century, researchers started to use non-acted speech [10, 11, 12]

INTRODUCTION

- According to Schlosberg [13], one can map different emotions as points of a two- or three-dimensional space



- Such an emotional space may be described in terms of **activation** and **potency**, among others.

- The so-called basic or archetypal emotions
  - happiness, anger, disgust, surprise, fear, and sadness: "Big Six" [14]

  lie in distinct "areas" of the emotional space.

- More recent studies suggest that emotional space is of even higher dimensionality [15]

INTRODUCTION

- Corpora

Characteristics of common emotional speech databases.

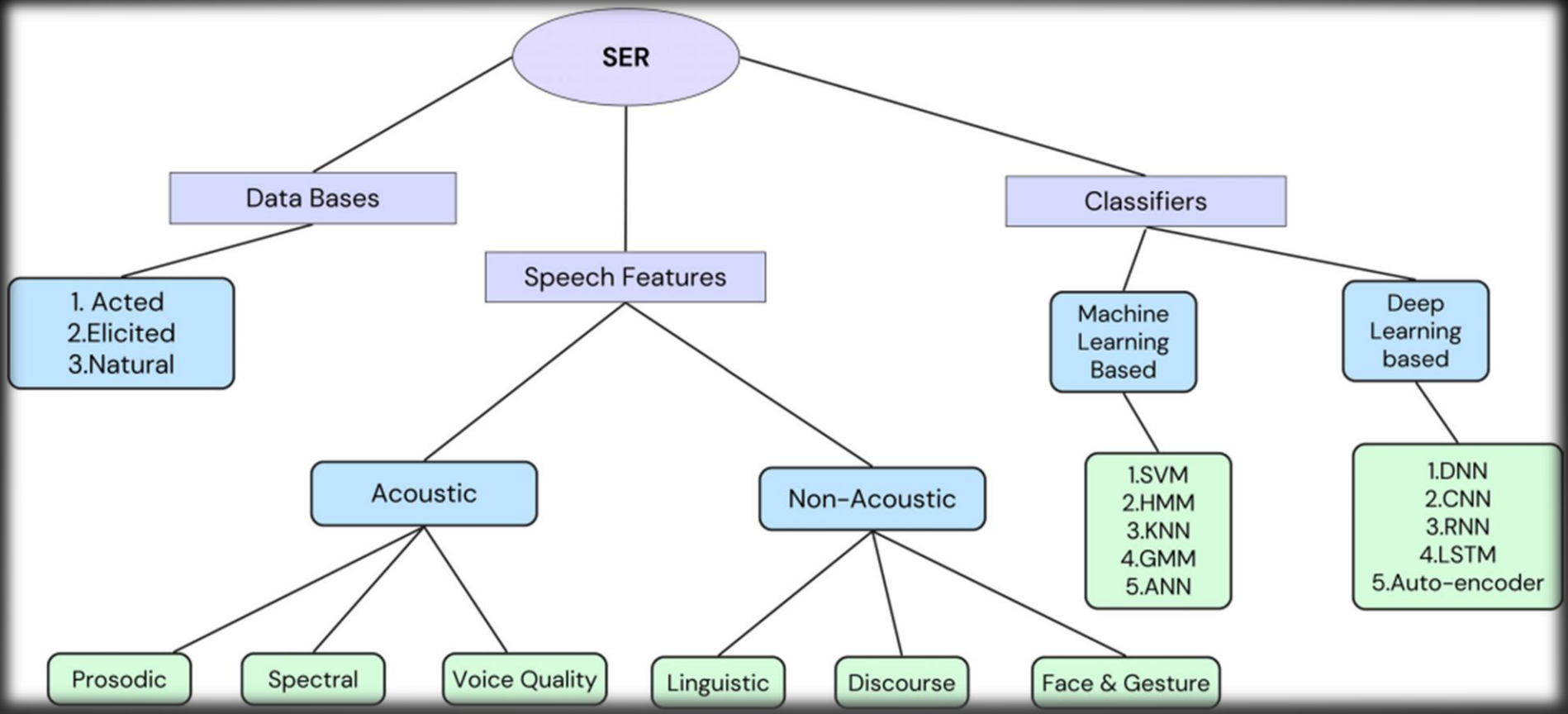
Corpus	Access	Language	Size	Source	Emotions
LDC Emotional Prosody Speech and Transcripts	Commercially available	English	7 actors × 15 emotions × 10 utterances	Professional actors	Neutral, panic, anxiety, hot anger, cold anger, despair, sadness, elation, joy, interest, boredom, shame, pride, contempt
Berlin emotional database	Public and free	German	800 utterances (10 actors × 7 emotions × 10 utterances + some second version) = 800 utterances	Professional actors	Anger, joy, sadness, fear, disgust, boredom, neutral
Danish emotional database	Public with license fee	Danish	4 actors × 5 emotions (2 words + 9 sentences + 2 passages)	Nonprofessional actors	Anger, joy, sadness, surprise, neutral
Natural	Private	Mandarin	388 utterances, 11 speakers, 2 emotions	Call centers	Anger, neutral
ESMBS	Private	Mandarin	720 utterances, 12 speakers, 6 emotions	Nonprofessional actors	Anger, joy, sadness, disgust, fear, surprise
INTERFACE	Commercially available	English, Slovenian, Spanish, French	English (186 utterances), Slovenian (190 utterances), Spanish (184 utterances), French (175 utterances)	Actors	Anger, disgust, fear, joy, surprise, sadness, slow neutral, fast neutral
KISMET	Private	American English	1002 utterances, 3 female speakers, 5 emotions	Nonprofessional actors	Approval, attention, prohibition, soothing, neutral
BabyEars	Private	English	509 utterances, 12 actors (6 males + 6 females), 3 emotions	Mothers and fathers	Approval, attention, prohibition
SUSAS	Public with license fee	English	16,000 utterances, 32 actors (13 females + 19 males)	Speech under simulated and actual stress	Four stress styles: Simulated Stress, Calibrated Workload Tracking Task, Acquisition and Compensatory Tracking Task, Amusement Park Roller-Coaster, Helicopter Cockpit Recordings
MPEG-4	Private	English	2440 utterances, 35 speakers	U.S. American movies	Joy, anger, disgust, fear, sadness, surprise, neutral
Beihang University	Private	Mandarin	7 actors × 5 emotions × 20 utterances	Nonprofessional actors	Anger, joy, sadness, disgust, surprise
FERMUS III	Public with license fee	German, English	2829 utterances, 7 emotions, 13 actors	Automotive environment	Anger, disgust, joy, neutral, sadness, surprise
KES	Private	Korean	5400 utterances, 10 actors	Nonprofessional actors	Neutral, joy, sadness, anger
CLDC	Private	Chinese	1200 utterances, 4 actors	Nonprofessional actors	Joy, anger, surprise, fear, neutral, sadness
Hao Hu et al.	Private	Chinese	8 actors × 5 emotions × 40 utterances	Nonprofessional actors	Anger, fear, joy, sadness, neutral
Amir et al.	Private	Hebrew	60 Hebrew and 1 Russian actors	Nonprofessional actors	Anger, disgust, fear, joy, neutral, sadness
Pereira	Private	English	2 actors × 5 emotions × 8 utterances	Nonprofessional actors	Hot anger, cold anger, joy, neutral, sadness

INTRODUCTION

- Corpora (more recent)

Database	Type	Modality	Size	Emotions	Language	Access Type	Purpose
Linguistic Data Consortium (Liberman et al. 2002) [56]	Acted	Audio	470 utterances	Hot anger, cold anger, Fear, disgust, Happiness, Sadness, Neutral, Panic, Pride, Despair, Interest, Elation, Shame, Boredom	English	Commercially available	Speech recognition, prosody, pronunciation modeling
BMS database new et al. (2003) [57]	Elicited	Audio	60 (Burmese) and 720 (Mandarin) utterances	Fear, anger, Disgust, Joy, Sadness, and Surprise	Burmese, Mandarin	Commercially available	Emotion recognition
SAVEE database (2014) [58]	Acted	Audio-visual	680 utterances	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral	English	Free to research use	Emotion recognition
GEMEP (2006) [59]	Acted	Audio-visual	7000+ utterances	Pride, Amusement, Elation, Interest, Pleasure, Relief, Various fears	German, French, Italian, English	Free for research purpose	Study of multi-modal expressions of emotion
MET database (2005) [36]	Acted	Audio	100 utterances	Neutral, Anger, Fear, Joy, Sadness, Disgust, Boredom	German	Free	Emotion recognition
RAVDESS (2018) [60]	Acted	Audio-Visual	7840 utterances	Calm, Happy, Sad, Angry, Fearful, Disgusted, Neutral, Surprised	English	Publicly accessible	Study of intensity of emotions in speech and song
TESS dataset (2019) [61]	Acted	Audio	2800 utterances	Anger, Disgust, Neutral, Fear, Happiness, Sadness, Pleasant, Surprise	English	Free	Psychological study of emotional speech
AFEW (2017) [62]	Natural	Audio-visual	1426 utterances	Anger, Sadness, Surprise, Disgust, Fear, Happiness, Neutral	English	Publicly accessible	Broad range of spontaneous emotion categories

INTRODUCTION

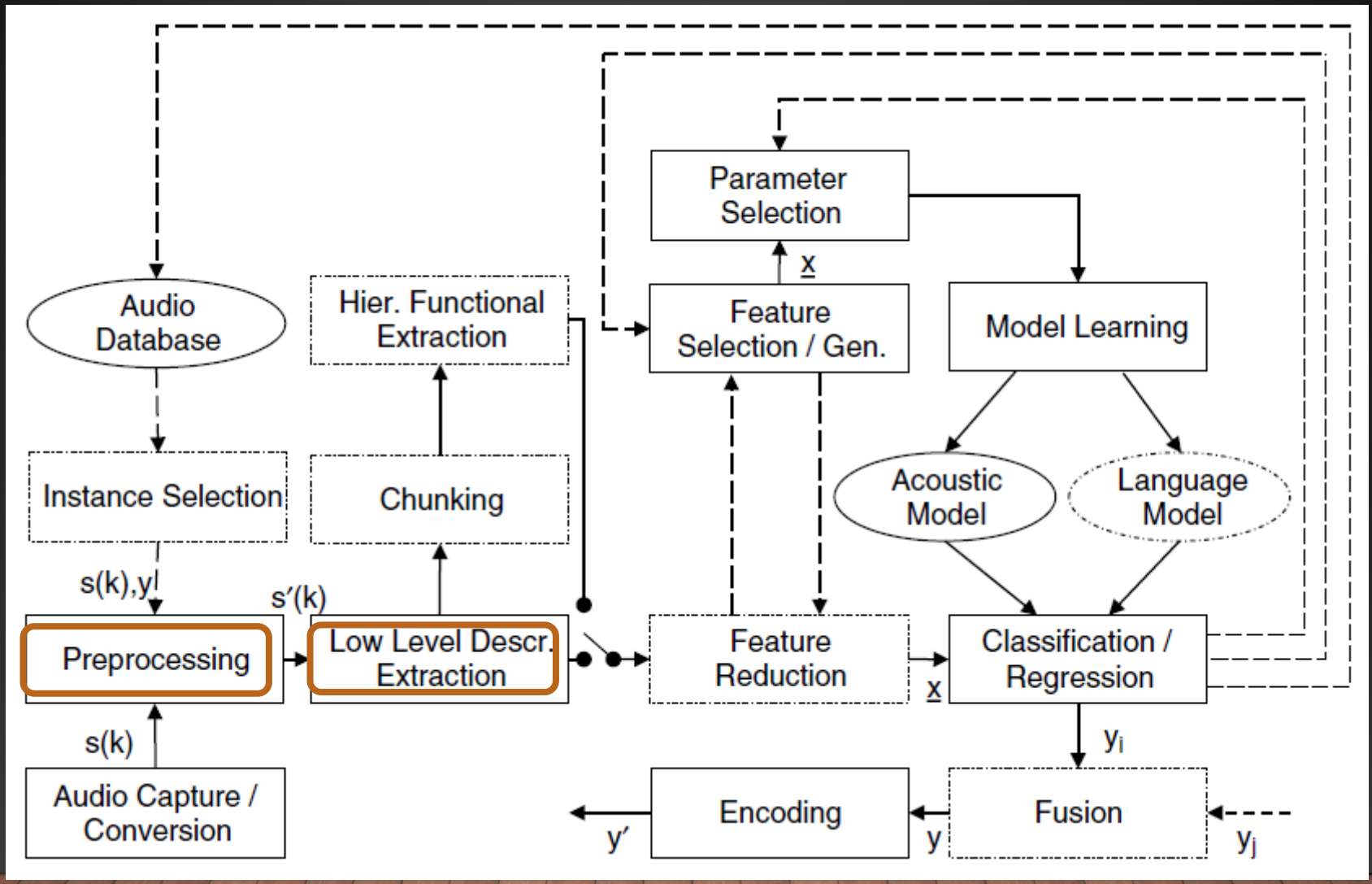


An overview of Speech Emotion Recognition [27].

OVERVIEW

- Introduction
- **Computational Modeling of Emotion**
- Acoustic Features
- Classifiers
- Conclusions

COMPUTATIONAL MODELING OF EMOTION



COMPUTATIONAL MODELING OF EMOTION

- Preprocessing
  - Goal: enhancement
    - Dereverberation
    - Denoising
    - Source separation
    - Automatic gain control
    - …and others
- Low-level Descriptor (LLD) Extraction:
  - Goal: extract features relevant to the emotional state
  - ~100 frames per second – 10-30 ms of window sizes
  - Time or frequency domain features

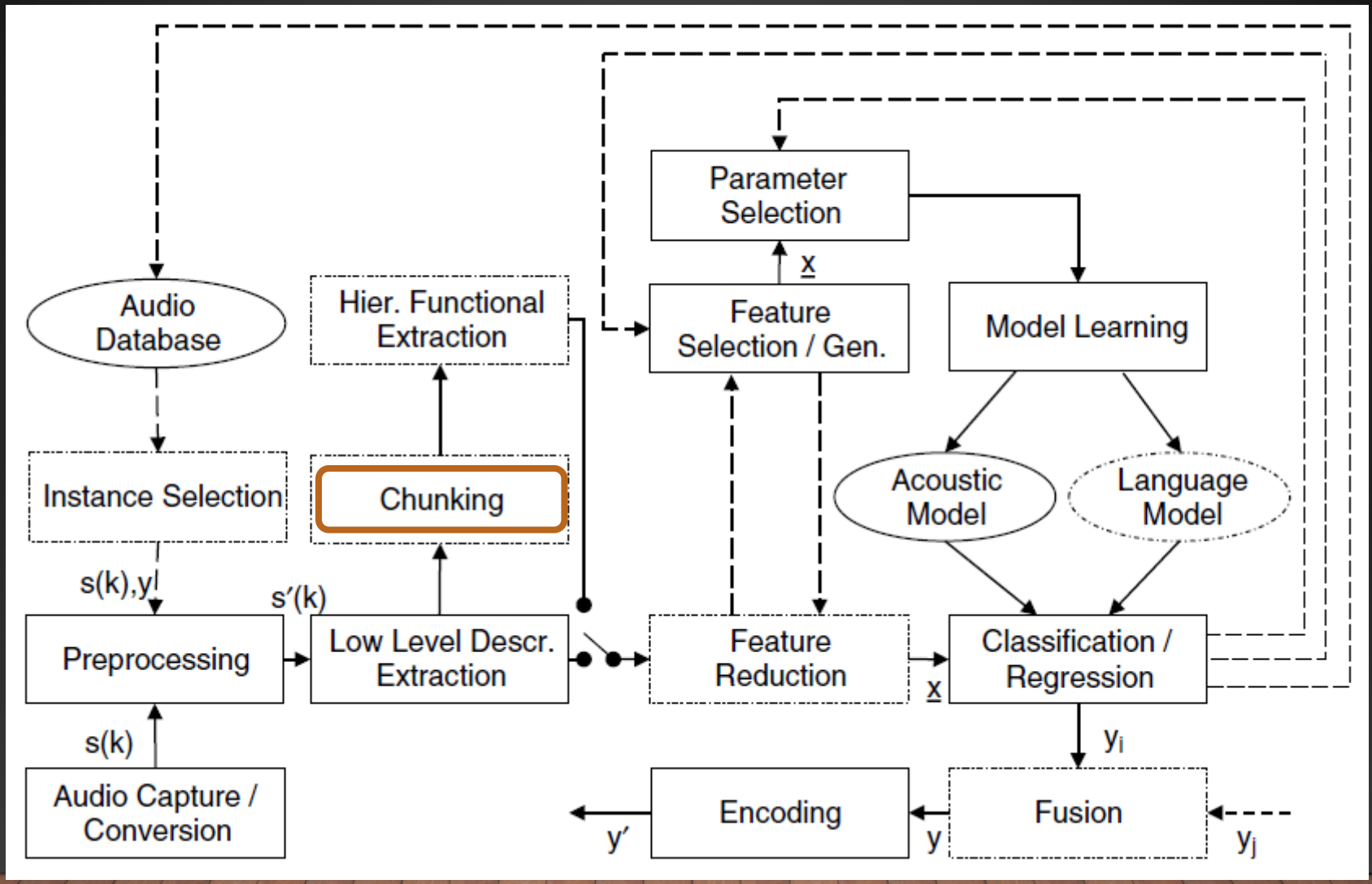
COMPUTATIONAL MODELING OF EMOTION

- Typical features:
  - Intensity (loudness, energy, etc.)
  - Intonation (pitch, etc.)
  - Linear Prediction Coefficients (LPCCs)
  - Perceptual Linear Prediction Coefficients (PLPCs)
  - Cepstral Coefficients (MFCCs, PLP-CCs, etc.)
  - Formants (amplitude, frequency, bandwidth, etc.)
  - Harmonicity (HNR, NHR, etc.)
  - Jitter and Shimmer
  - Spectral Statistics (flux, variance, slope, etc.)
  - … plus many more!

COMPUTATIONAL MODELING OF EMOTION

- These features can be augmented by further descriptors such as Deltas (Δ) and Delta-Deltas (ΔΔ), regression coefficients, etc.
- Moreover, smoothing and mean/variance normalization are quite often in practice.
- Linguistic LLDs
  - Phoneme sequences, word sequences
  - Laughter, sighs
  - Pauses, hesitations
  - …and others
- Their extraction often requires an ASR system

COMPUTATIONAL MODELING OF EMOTION

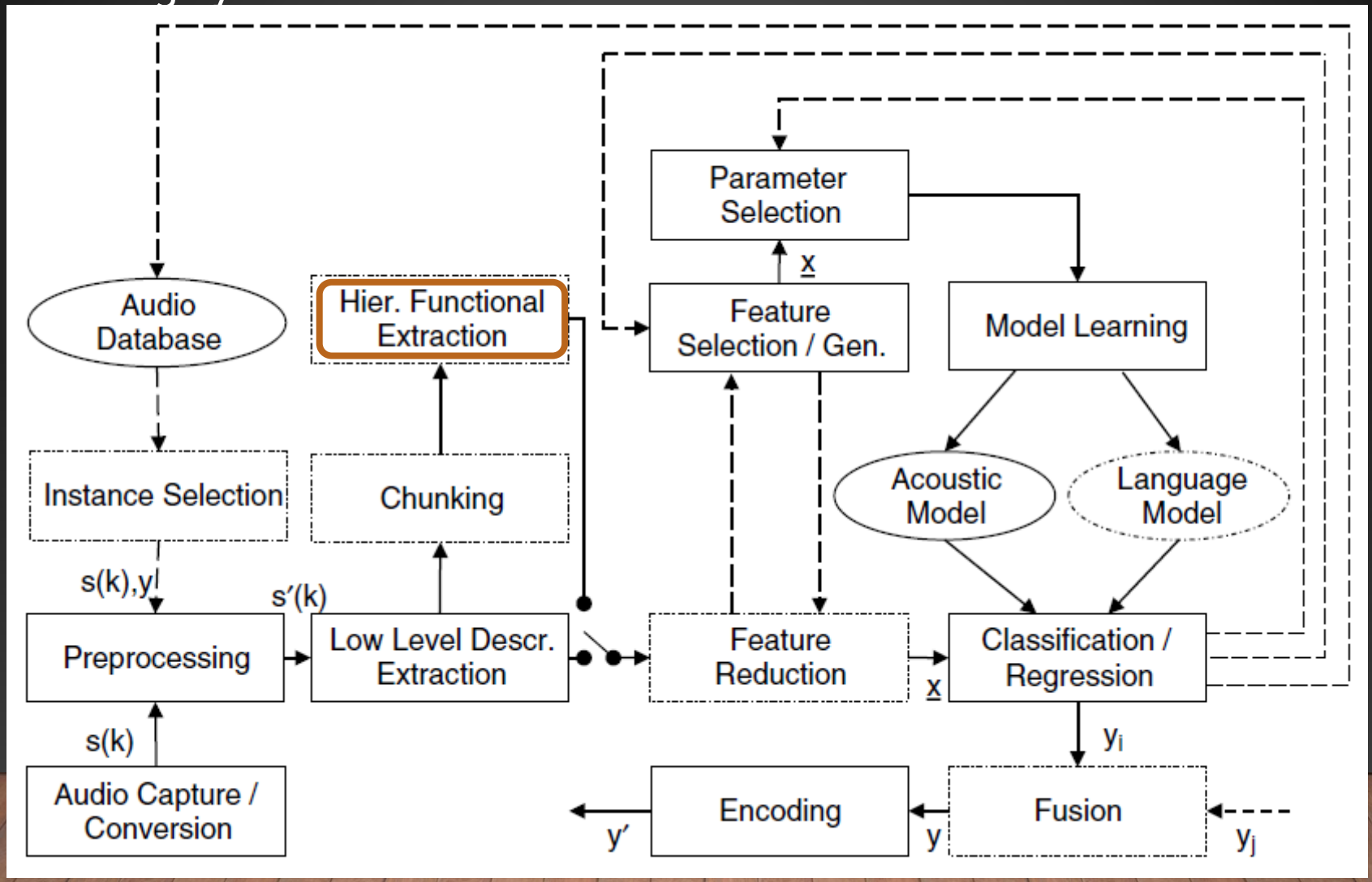


COMPUTATIONAL MODELING OF EMOTION

- Chunking (opt.)
  - Most phenomena in emotion analysis are expressed by the evolution of certain LLDs over time
  - Group LLD frames into meaningful temporal units of analysis
  - Hundreds of ms < unit length < a few seconds
  - Types of units investigated:
    - Fixed number of frames
    - Acoustic chunking
    - Voiced/unvoiced parts of speech
    - Phonemes
    - Syllables
    - Words
    - Even complete sentences!
  - Such tasks require pre-analysis
    - Speech activity detection, V/UV detection, structural analysis, etc.

COMPUTATIONAL MODELING OF EMOTION

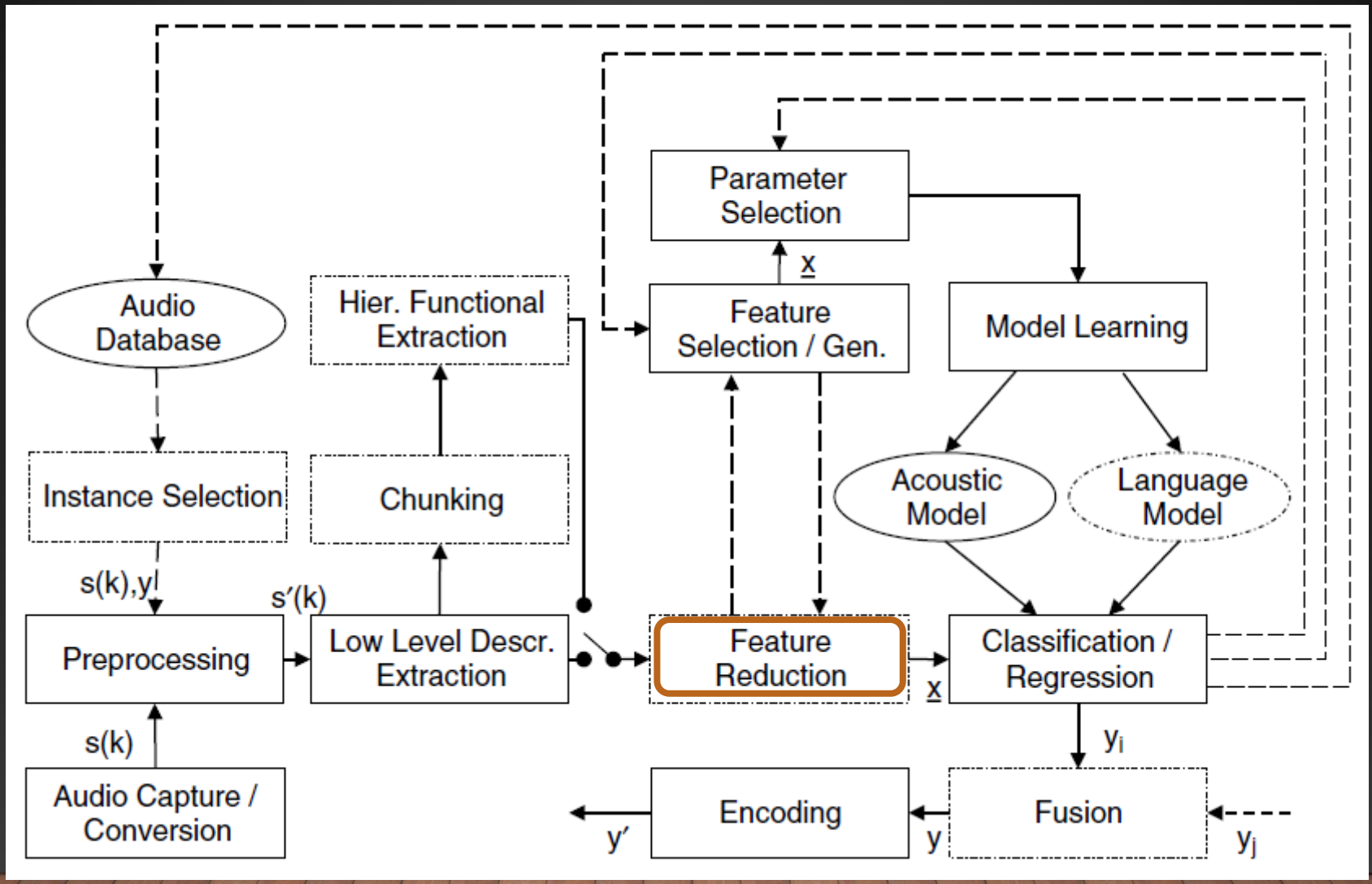
- Let's now discuss the main building blocks of an emotional state modeling system



COMPUTATIONAL MODELING OF EMOTION

- Hierarchical Functional Extraction (opt.)
  - Functionals are applied to each LLD within the analysis window [16, 17]
  - Why?
    - To further reduce information
    - To project the LLD time series of potentially unknown and variable length to a scalar value for each applied functional and LLD.
  - This is called "supra-segmental" analysis.
  - Some well-known functionals for speech chunks are:
    - Extremes (min, max, range, etc.)
    - Means (arithmetic, absolute, etc.)
    - Percentiles (quartiles, ranges, etc.)
    - Standard deviation
    - Higher moments (kyrtosis, skewness, etc.)
    - Tempo (duration, positions, etc.)
  - Also, there are linguistic functionals

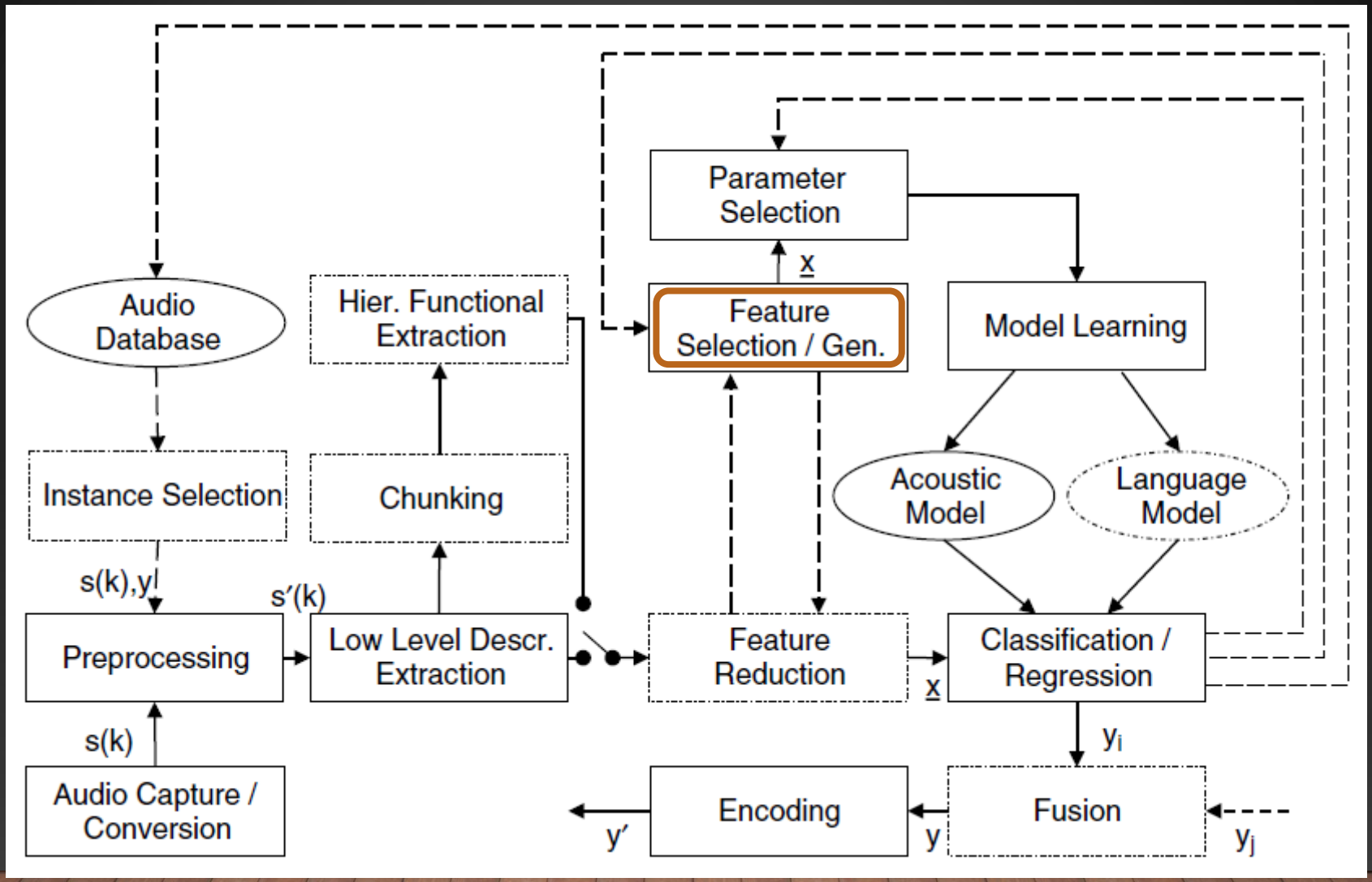
COMPUTATIONAL MODELING OF EMOTION



COMPUTATIONAL MODELING OF EMOTION

- Feature reduction
  - Goal: remove redundant information from features **and** keep information related to the target of interest
  - Transformation of the feature space
    - Translation into the origin of the space
    - Rotation to reduce covariance between features in the transformed space
  - Usual algorithms:
    - Principal Component Analysis (PCA)
    - Linear Discriminant Analysis (LDA)

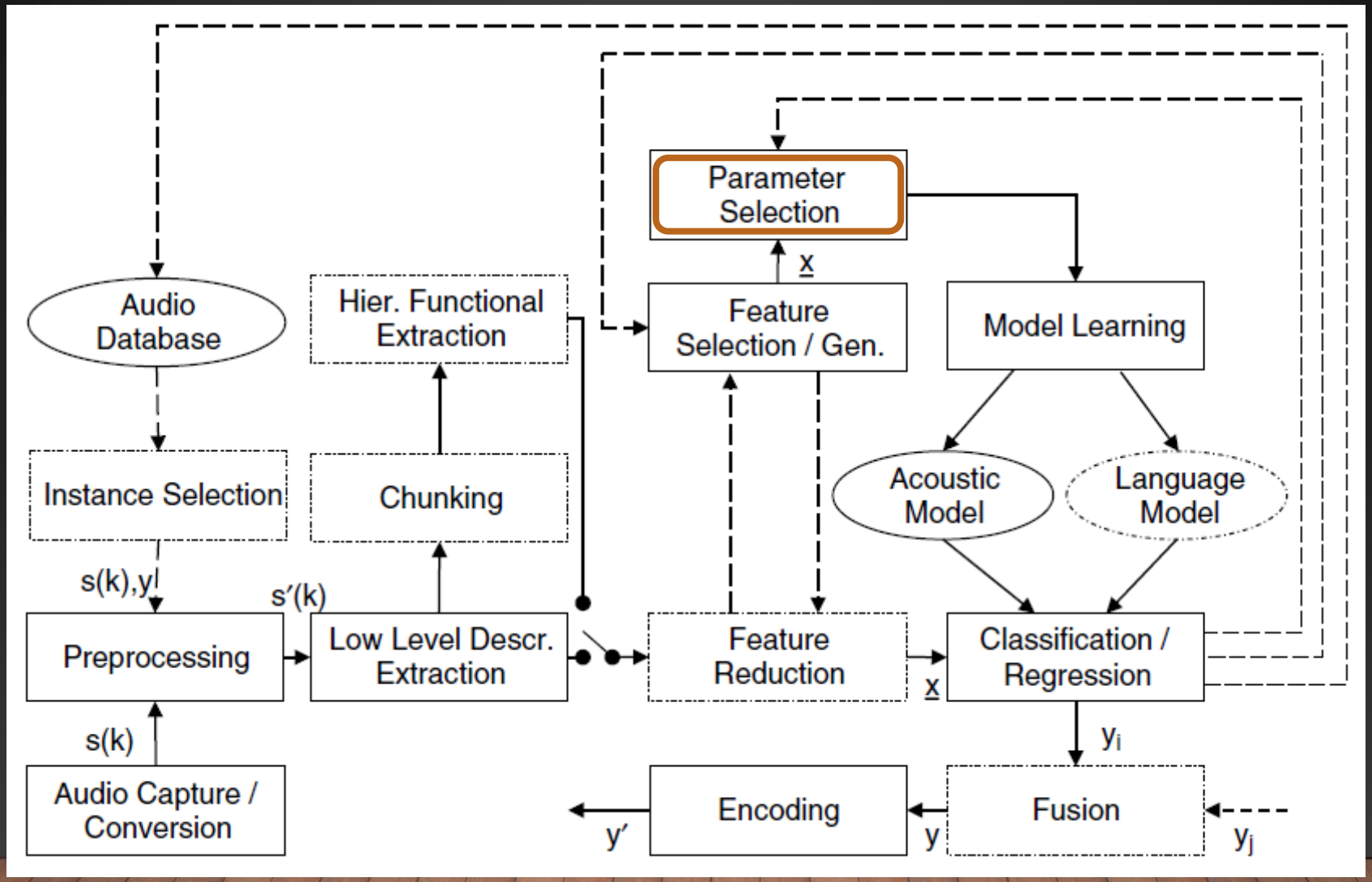
COMPUTATIONAL MODELING OF EMOTION



COMPUTATIONAL MODELING OF EMOTION

- Feature selection
  - Goal: find the "best" subset of features according to a predefined selection criterion
  - Reasons:
    - Features can be:
      - Irrelevant (no effect on processing)
      - Redundant (same, highly correlated)
    - Decrease problem dimensionality
  - We keep only those who are well-correlated with the task at hand
    - A target function is decided at first step
    - Open loop selection: select features for which the reduced data maximizes between-class separability
    - Closed loop selection: select features based on the processing algorithm's performance that serves as a criterion for a feature subset selection

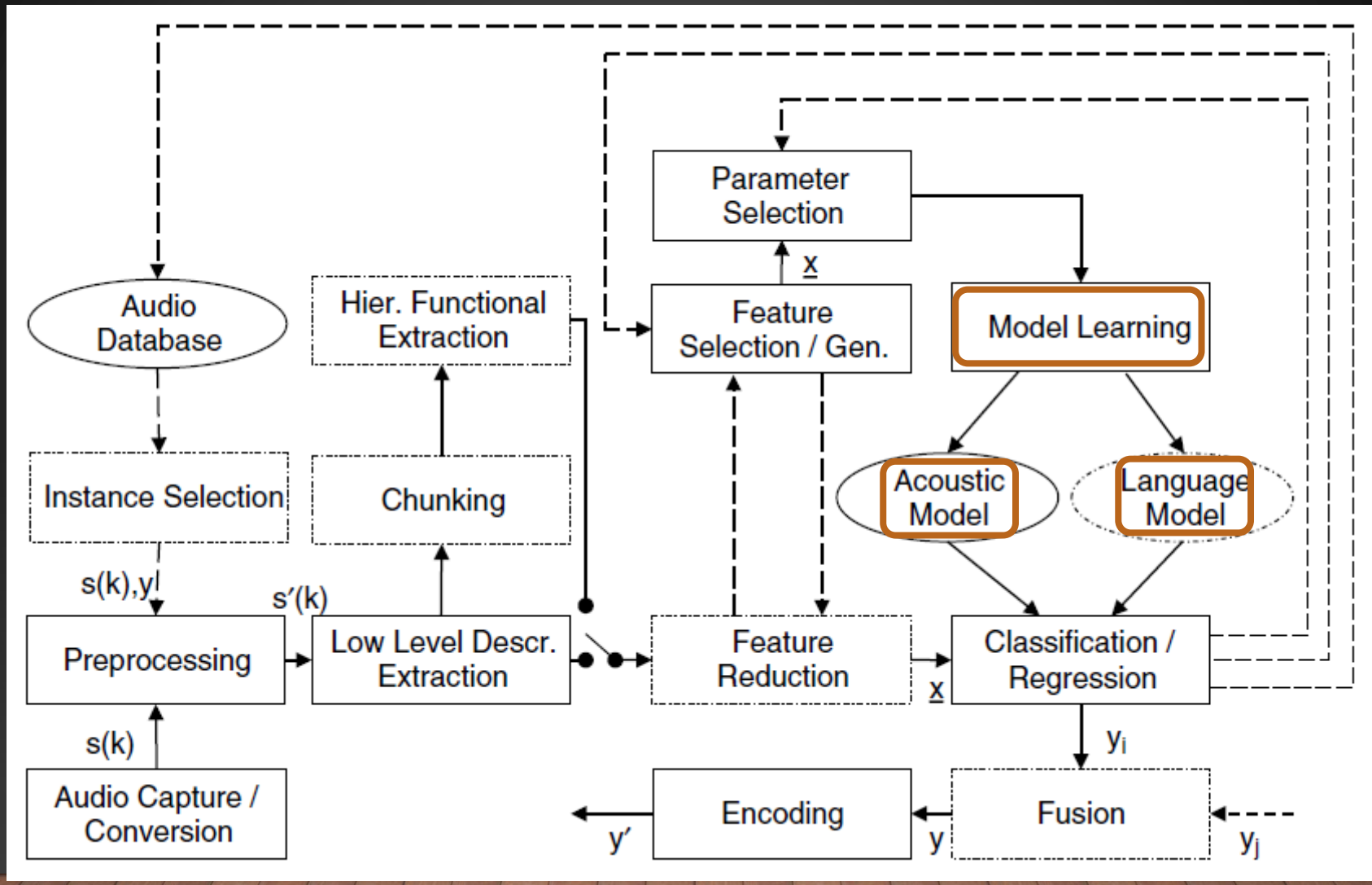
COMPUTATIONAL MODELING OF EMOTION



COMPUTATIONAL MODELING OF EMOTION

- Parameter Selection
  - Fine-tuning of the learning algorithm and the models
    - Optimization of a model's topology
    - Initialization values
    - Type of kernel functions used
    - Step sizes
    - Number of iterations in the learning phase
  - Grid search is the most common approach

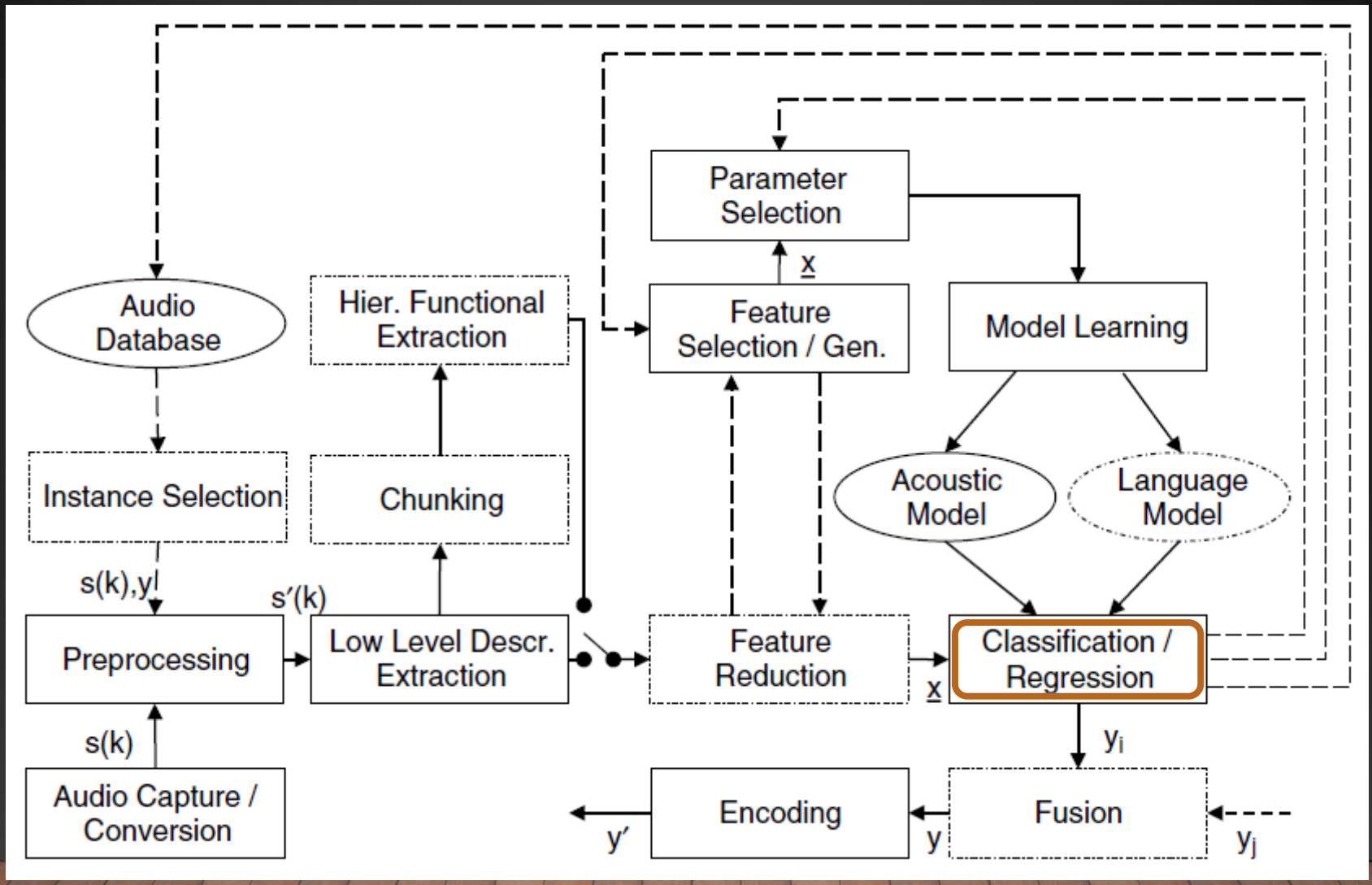
COMPUTATIONAL MODELING OF EMOTION



COMPUTATIONAL MODELING OF EMOTION

- Model learning
  - Supervised learning
    - Model is built based on labelled data
- Acoustic Model
  - It contains a numerical representation of the learnt dependencies between the acoustic observations and the labels (classes)
- Language Model (opt.)
  - It contains statistical dependencies between units of the input language (words, syllables, etc.)

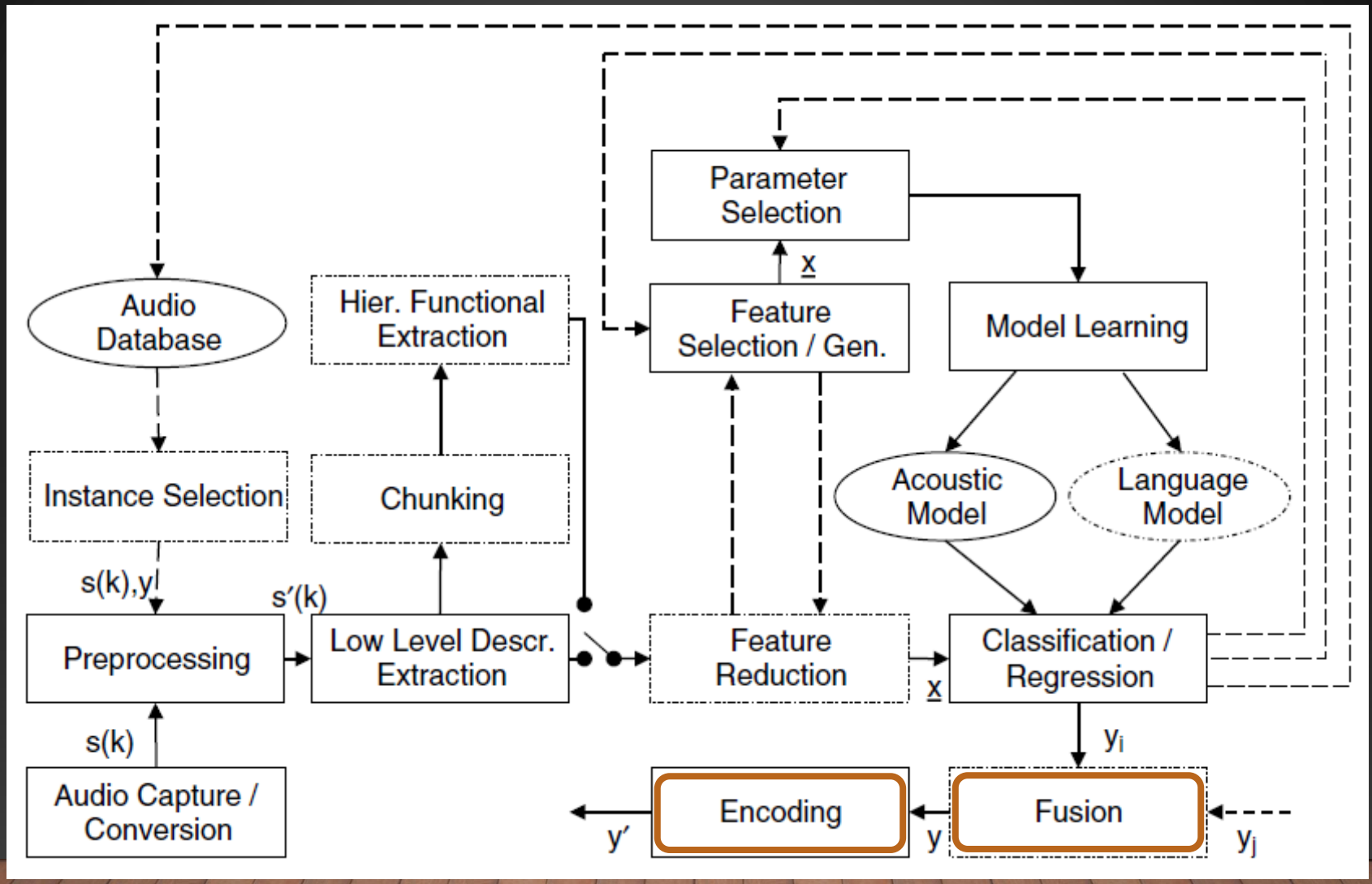
COMPUTATIONAL MODELING OF EMOTION



COMPUTATIONAL MODELING OF EMOTION

- Classification/regression
  - A target label is assigned to an unknown test instance
  - Classification: discrete labels
  - Regression: continuous values
  - High diversity of classifiers and regressors in the field

COMPUTATIONAL MODELING OF EMOTION



COMPUTATIONAL MODELING OF EMOTION

- Fusion (opt.)
  - Stage where information from different input streams is to be fused at the "late semantic" level (fusion of labels and scores)
- Enconding (opt.)
  - Once the final label has been assigned, the information needs to be represented in an optimal way for further processing by other systems (e.g. spoken dialogue systems)
    - Additional information such as confidence scores can be encoded
    - Standards: VoiceXML, Emotion Markup Language (EML), etc.

OVERVIEW

- Introduction
- Computational Modeling of Emotion
- **Acoustic Features**
- Classifiers
- Conclusions

ACOUSTIC FEATURES

- We will present some LLDs that are frequently used in the literature
- Short time analysis of speech [18]
  - Speech are non stationary signals
  - We need to process them in short segments (frames) that are approximately stationary
  - Windowing
    - Long enough to reliably estimate feature
    - Short enough to ensure the feature does not change inside the window
    - Trade-off: 20-40 ms is a good compromise
    - Pitch synchronous or asynchronous approaches
    - Type of window: Hamming, Hanning, Rectangular, Bartlett, Kaiser, etc.
  - Step size
    - Usually equal to 10 ms → 50% overlap between frames
    - "How often you sample the speech signal"

ACOUSTIC FEATURES

- Continuous Descriptors
  - Acoustic LLDs or Acoustic Features
  - We will discuss
    - Intensity
    - Fundamental frequency F0
    - Voicing probability
    - Formants & Bandwidths
    - Jitter & Shimmer
    - …and many more! ☺

ACOUSTIC FEATURES

- Intensity
  - Intensity is related to pitch, duration, and spectral shape of stimulus
  - Instead we use the **Energy** of the signal,

$$E = \sum_{k=-\infty}^{\infty} s^2[k]$$

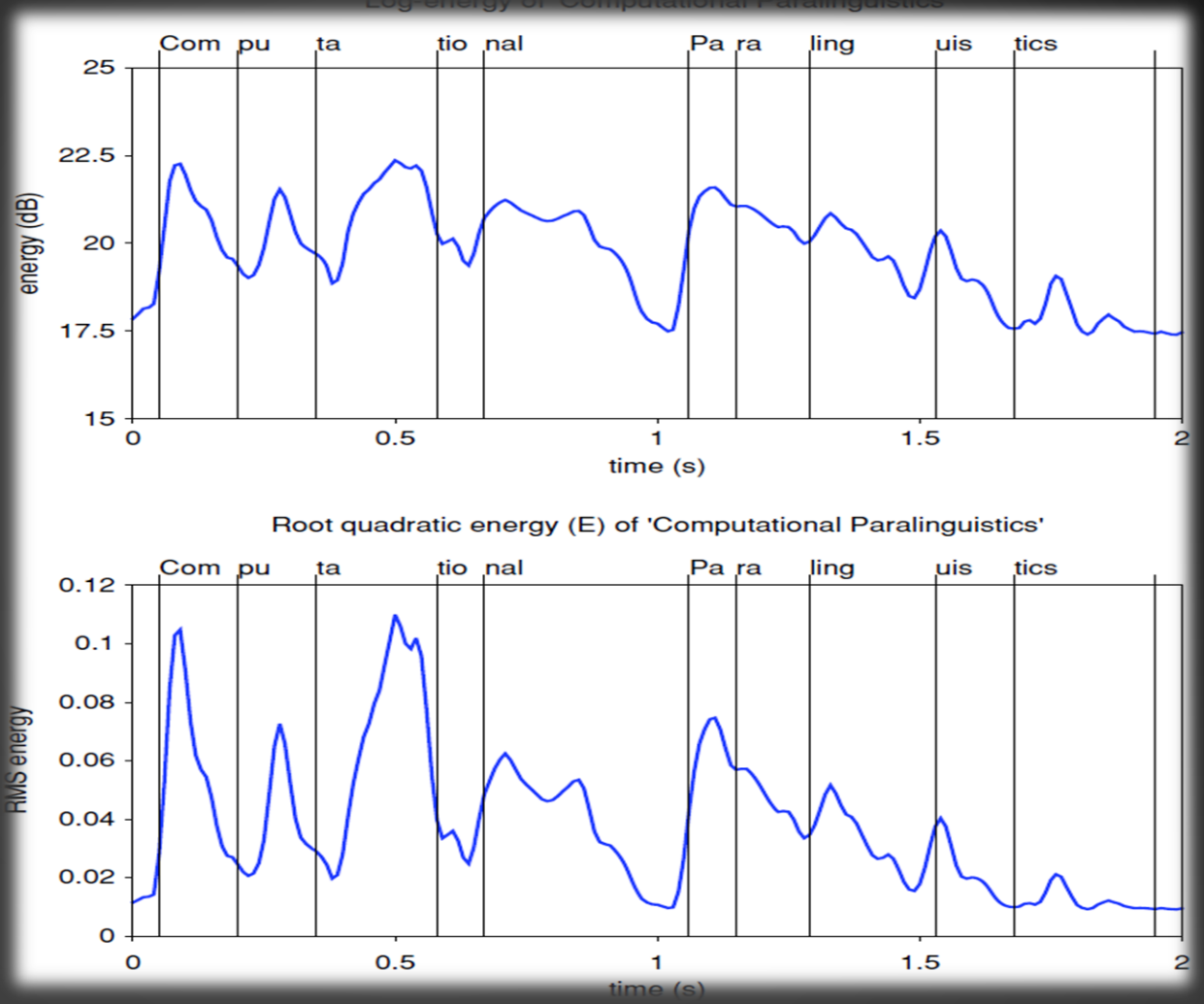
  - For a speech frame at time instant n:

$$E(n) = \sum_{k=-\infty}^{\infty} (s[k]w[n-k])^2$$

    where $w[n]$ is the window function supported in $[n-M, n+M]$.
  - Alternatively, we can use the **Root Mean Square Energy**
    - It's just the square root of the energy and it's numerically more constrained
  - Or the **Log-Energy** (just the logarithm of energy)
    - Closer to how we perceive loudness

ACOUSTIC FEATURES

- Intensity



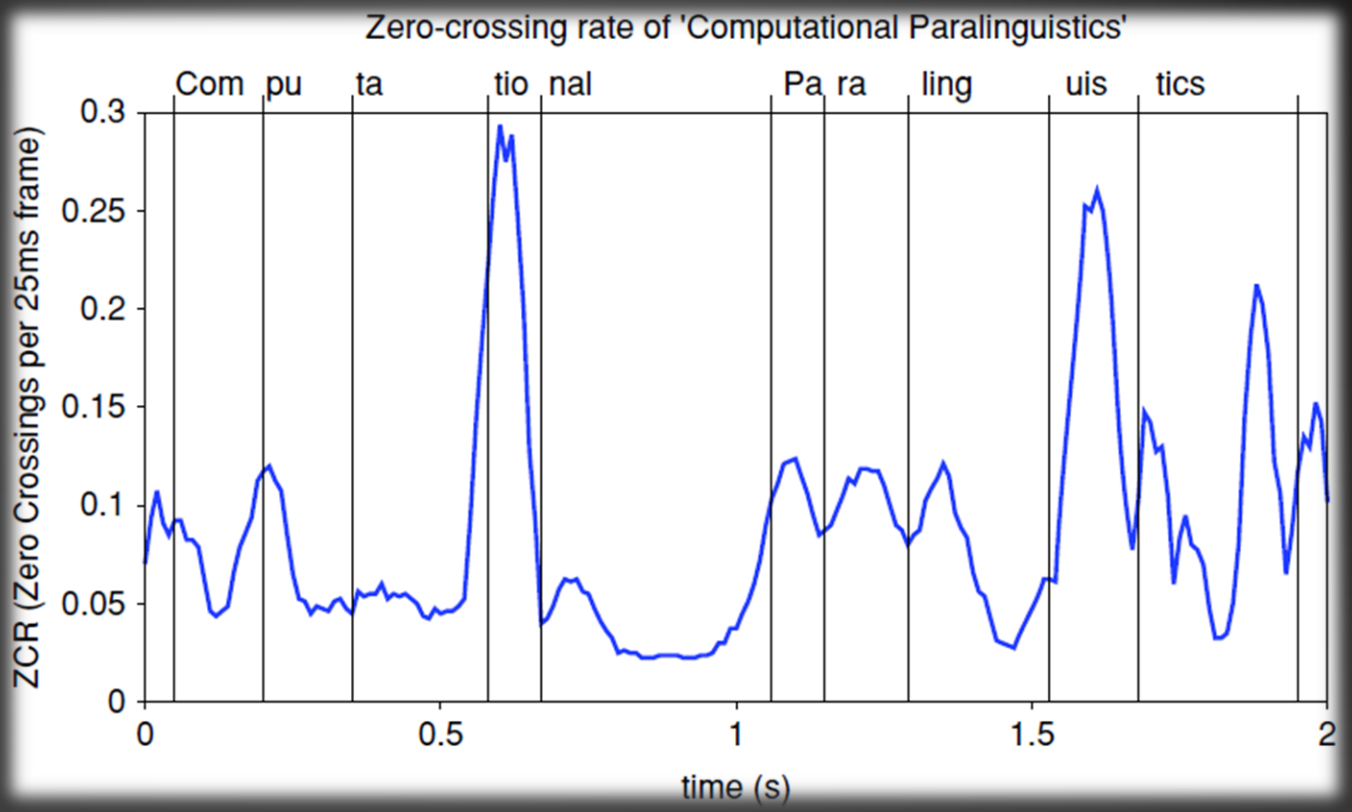
ACOUSTIC FEATURES

- Zero Crossings Rate (ZCR)
  - The ZCR provides information about the frequency distribution [19]
  - High ZCR → speech has high frequency components
  - Low ZCR → speech has low frequency components
  - Useful for voiced/unvoiced discrimination
  - Intuition:
    - For a pure sinusoid, the ZCR is twice its frequency
  - ZCR is given by

$$ZCR(n) = \sum_{k=-\infty}^{\infty} \frac{1}{2} |sgn(x[k]) - sgn(x[k-1])| w[n-k]$$

ACOUSTIC FEATURES

- Zero Crossings Rate (ZCR)



ACOUSTIC FEATURES

- Autocorrelation function (ACF)
  - General definition

$$R(d) = \sum_{k=-\infty}^{\infty} s[k]s[k+d]$$

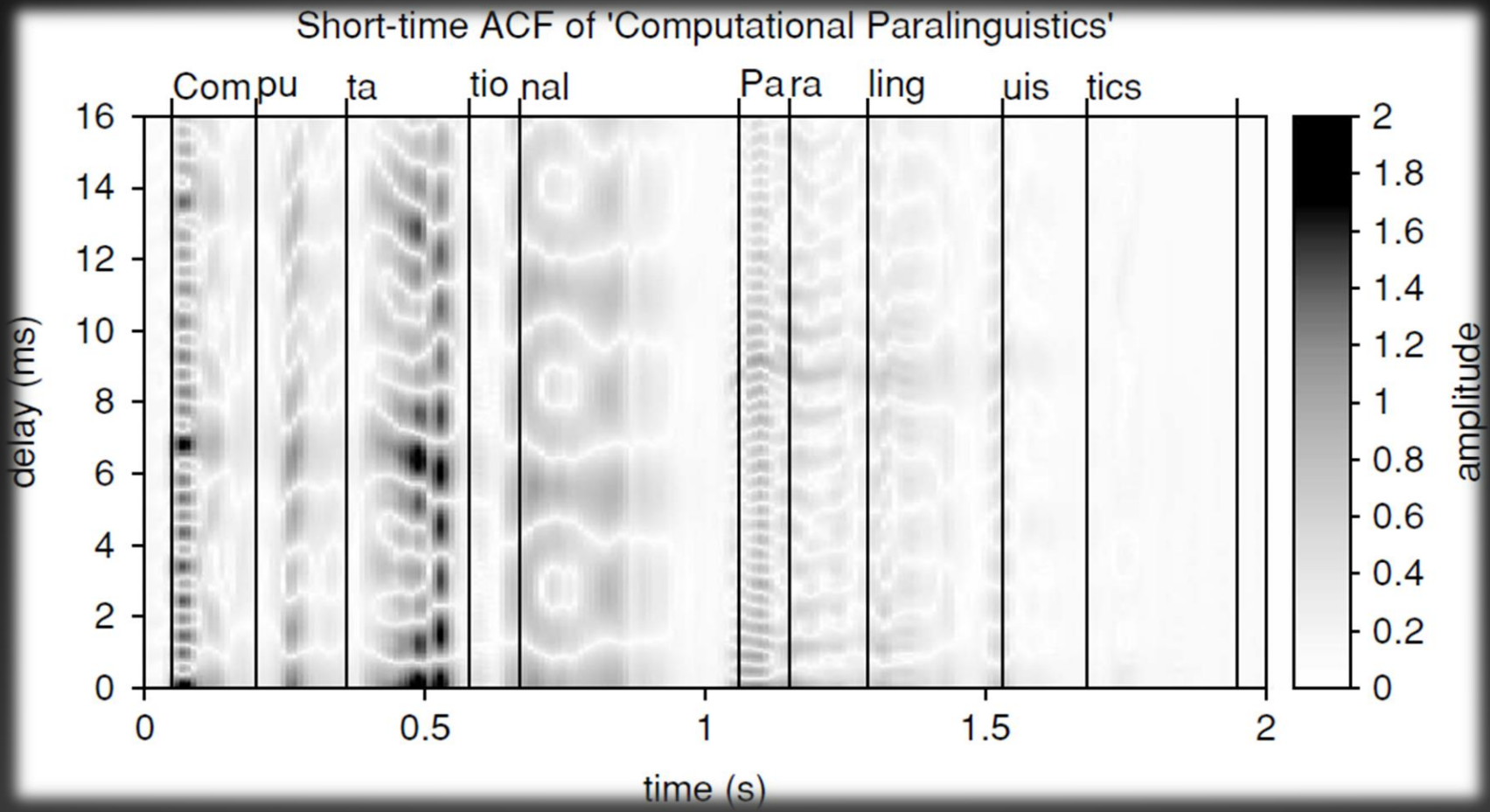
  - Short time autocorrelation

$$R(n,l) = \sum_{k=-\infty}^{\infty} s[k]s[k+l]w[n-k]w[n-k+l]$$

  - Properties [20]:
    - $R(0)$ is equal to the signal's energy
    - Periodic if signal is periodic
    - Scaling the signal by a results in scaling the ACF by a^2

ACOUSTIC FEATURES

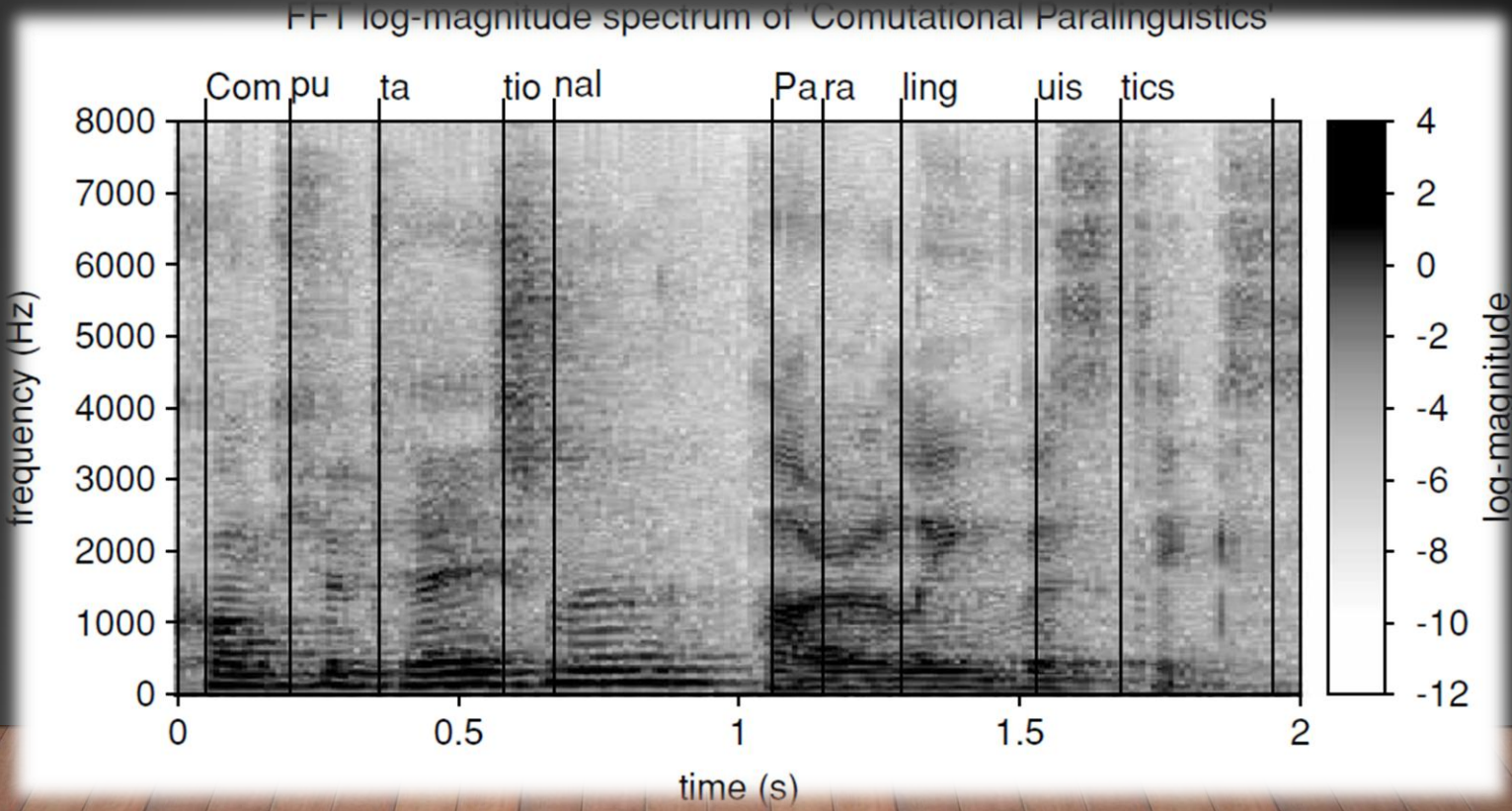
- Autocorrelation



ACOUSTIC FEATURES

- Spectrum
  - Short time spectrum [21]

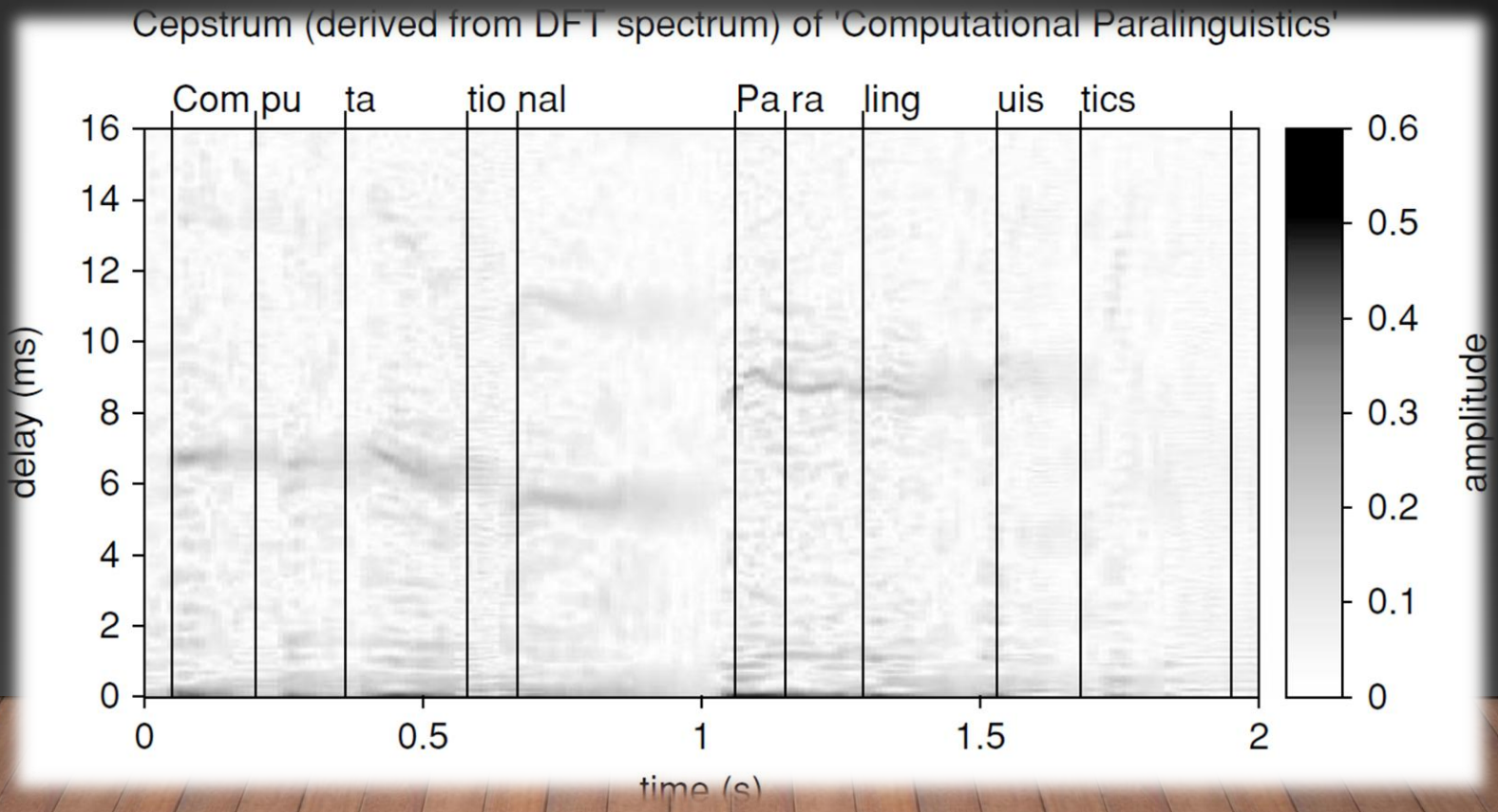
$$S(n,\omega) = \sum_{m=-\infty}^{\infty} s[m]w[n-m]e^{-j\omega m}$$



ACOUSTIC FEATURES

- Cepstrum
  - The cepstrum is the Inverse F.T. of the log-magnitude spectrum [18]
  - Short time cepstrum

$$c(k,n) = \mathrm{IFT}(\log|S(n,\omega)|)$$



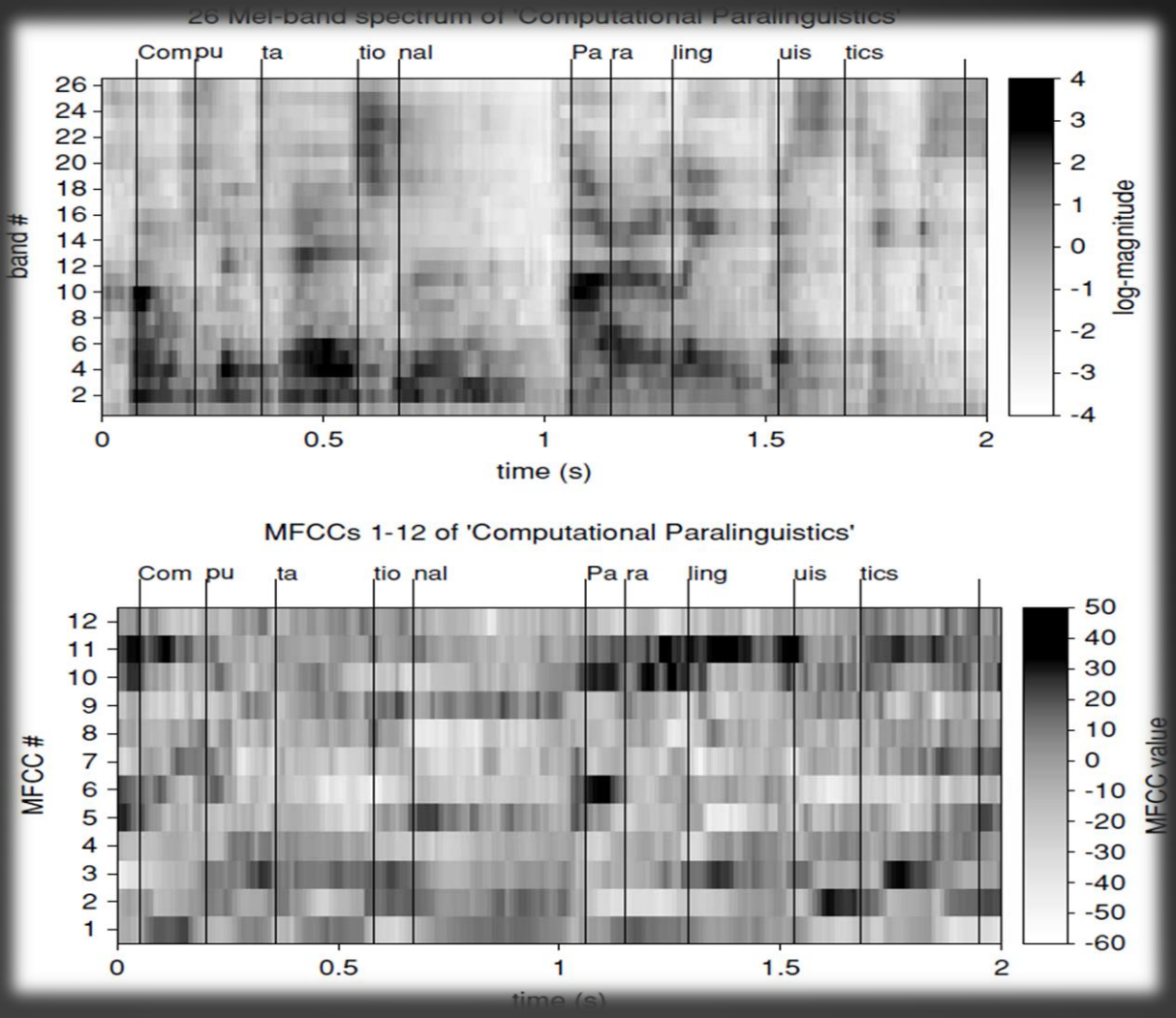
ACOUSTIC FEATURES

- Mel Frequency Cepstral Coefficients (MFCCs)
  - Goal: take human perception into account [22]
    - Map the power spectrum onto Mel-scale bands
      - …using triangular, equidistant, overlapping windows
    - Take the logarithm in each frequency band
    - Perform a Discrete Cosine Transform (DCT)
  - In mel-scale, the lower frequencies are better resolved
    - …thus mimicking human hearing
  - Conversion from Hz to Mel scale

$$\mathrm{Mel}(f) = 2595 \log\left(1 + \frac{f}{700}\right)$$

ACOUSTIC FEATURES

- MFCCs



ACOUSTIC FEATURES

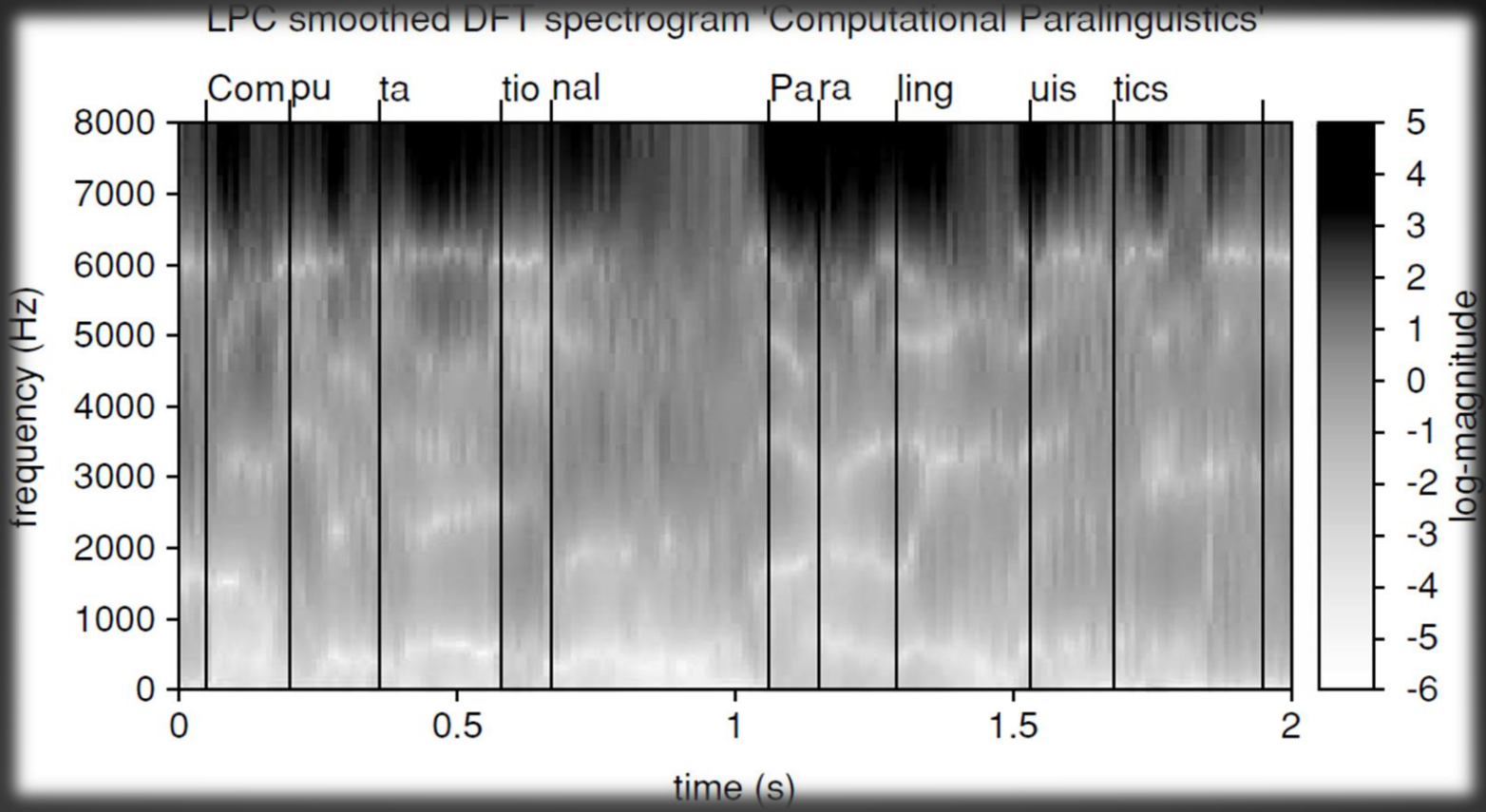
- Linear Prediction Coefficients (LPCCs)
  - Goal: voiced sounds are well-modeled by the vocal tract resonant frequencies: the *formants*
    - This means that consecutive voiced speech samples are correlated
      - Linear dependencies exist among them
      - We can predict a sample from the previous ones
  - Linear Prediction [18]

$$\hat{s}[n] = -\sum_{i=1}^{p} a_i s[n-i]$$

  - p is the order of the predictor
  - a_is are the prediction coefficients
    - They should be predicted in short time segments (frames) due to the non-stationarity of the speech waveform
    - Levinson-Durbin recursion is a well-known algorithm

ACOUSTIC FEATURES

- LPCCs

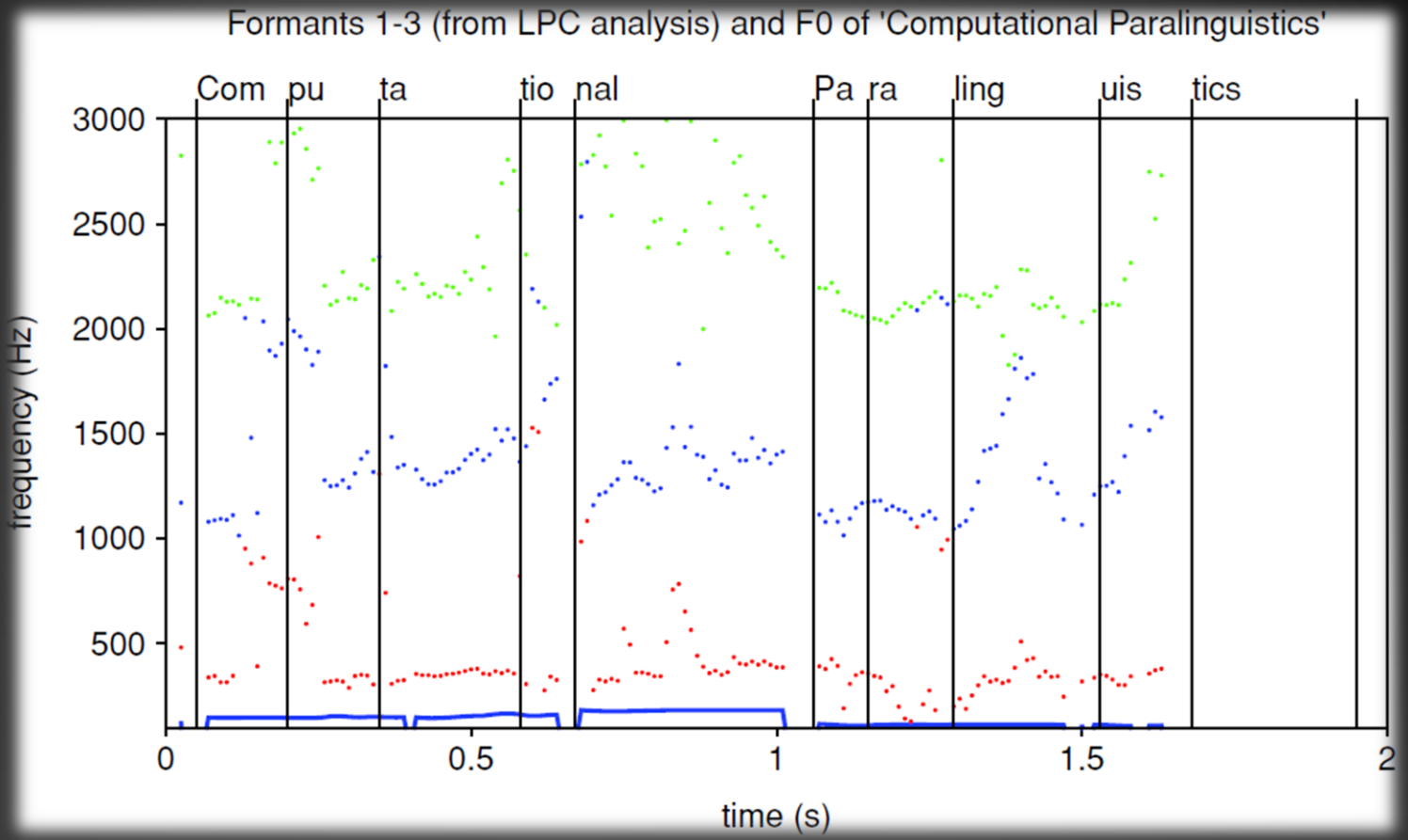


ACOUSTIC FEATURES

- Formants
  - Resonance frequencies of the vocal tract transfer function
  - F1 and F2 correlate well with the phonetic content
    - Higher formants describe speaker characteristics
  - Consist of:
    - Center frequency
    - Bandwidth
    - Amplitude
  - Usually, LPC is used for extraction of formants
    - However, difficulties arise when we use a spectral representation for formants
      - Spurious spikes
      - Limited spectral resolution
      - Nasality (introduction of new formants – compensation by zeros)
  - It is not considered as a solved problem ☺

ACOUSTIC FEATURES

- Formants

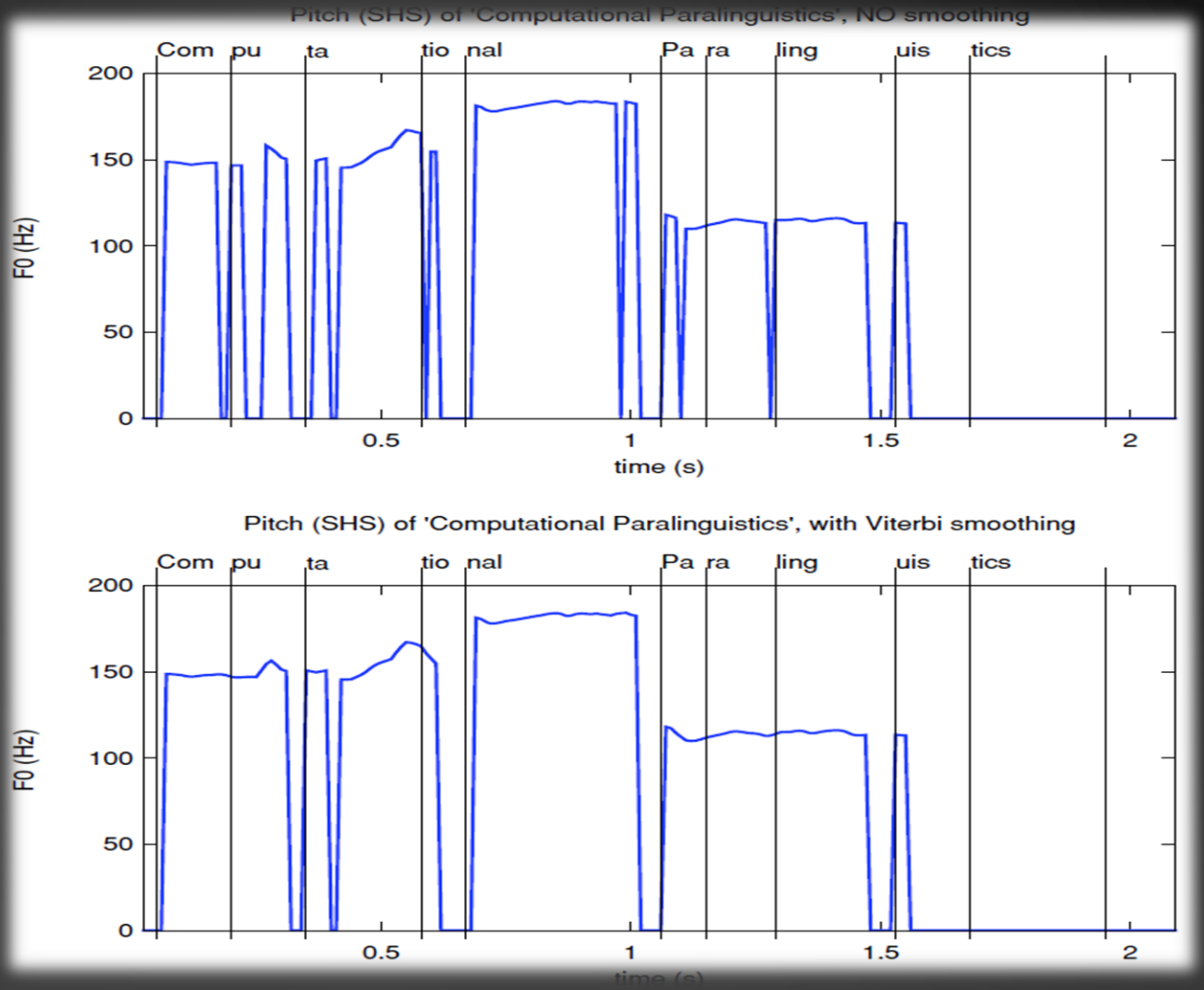


ACOUSTIC FEATURES

- Fundamental frequency
  - Human perception is far more sensitive to changes in the fundamental frequency than any other speech parameter! [25]
    - Thus, it has a significant influence on the performance of algorithms for emotion detection in speech
    - It is one of the most difficult tasks in speech processing!
      - Non-stationarity of speech
      - Confusion with low formants
      - Periodicity is not always regular (think pathological voices)
      - Range: [50-800] Hz (including children)
      - Missing fundamental
  - Algorithms aim for
    - F0 (average over a short analysis window)
    - T0 = 1/F0 (momentary value)
  - Steps: pre-process, extraction, post-process

ACOUSTIC FEATURES

- Fundamental frequency



ACOUSTIC FEATURES

- Jitter and Shimmer
  - Voice quality features
  - Represent micro-perturbations in frequency and amplitude of the excitation signal
    - They are called micro-prosodic descriptors
  - Jitter: deviation of the length of the fundamental period from one period to the next
    - Speaker age, voice pathology determination, etc.
    - Local jitter:

$$J_{pp} = T_0(n) - T_0(n-1)$$

    - Cycle jitter:

$$J_c = T_0(n) - \widehat{T_0}$$

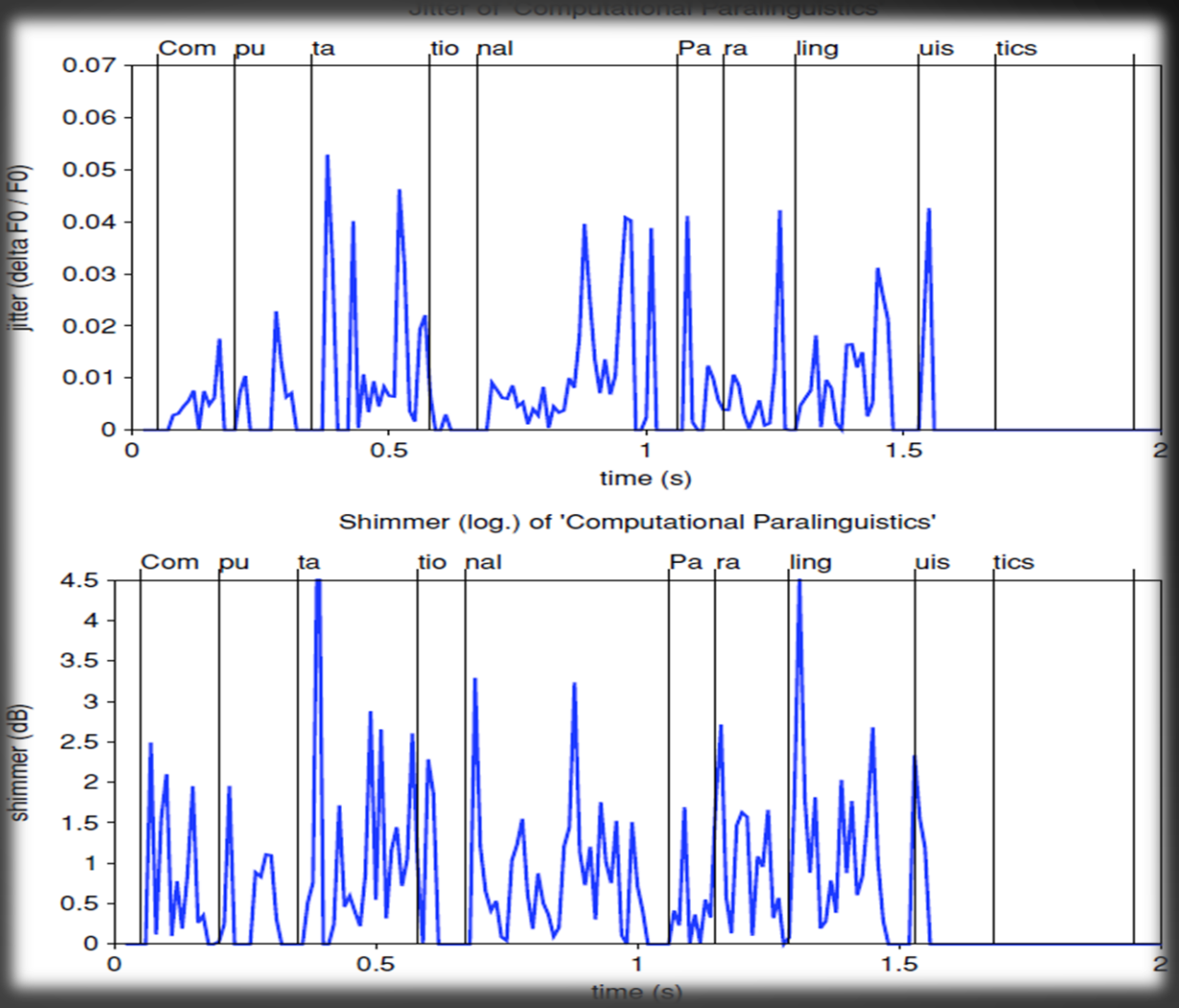
    where $\widehat{T_0}$ is the "ideal" fundamental period estimated as an average of all pitch periods in the analysis interval
  - Jitter is known to be high at the beginning and end of a sustained vowel

ACOUSTIC FEATURES

- Jitter and Shimmer
  - Shimmer: variation of the amplitude from one period to the next
    - It's measured on a log-scale (dB)
    - Healthy person's shimmer: 0.05-0.22 dB
  - Over successive speech cycles, jitter and shimmer help to give the vowel its naturalness
  - Moreover, the two phenomena contribute to the voice quality of a speaker.
  - A high degree of jitter and/or shimmer results in a voice with roughness, that is usually perceived as noise in recordings of pathological voices.
  - The actual measurements of jitter and shimmer may take place in
    - the time domain,
    - the frequency domain (magnitude spectrum), or
    - the quefrency domain

ACOUSTIC FEATURES

- Jitter and Shimmer



ACOUSTIC FEATURES

- Derived Low-level Descriptors
  - Derived features can be computed from the just discussed features [17]
    - Little research has been done towards this direction but it seems promising!
  - Two well-known methods:
    - Delta regression coefficients (first differential)

$$d(t) = \frac{\sum_{i=1}^{N} i(x(t+i) - x(t-i))}{2\sum_{i=1}^{N} i^2}$$

    - Delta-delta coefficients are very common as well
      - Apply same equation but on d(t) this time
    - Smoothing of LLDs

$$\hat{x}(t) = \frac{1}{N}\sum_{i=0}^{N-1} x\left(t - \frac{N-1}{2} + i\right)$$

with $N > 2$ and odd

OVERVIEW

- Introduction
- Computational Modeling of Emotion
- Acoustic Features
- **Classifiers**
- Conclusions

CLASSIFIERS

- For possible classification algorithms, the list is endless… ☺
- Classical ML:
  - Support Vector Machines, Decision Trees, Random Forests, K-Nearest Neighbors, Hidden Markov Models, Gaussian Mixture Models, Boosting Ensembles, etc
- Modern ML: Deep Neural Networks
  - Convolutional Neural Networks, (Bi-)LSTM Networks, Residual Networks, Capsule Networks, Graph Neural Networks, Pre-trained models, Transformers……..
- …and maaaaany more ☺

CLASSIFIERS

- Classification performance as reported in literature [26]:

Classification performance of popular classifiers, employed for the task of speech emotion recognition.

Classifier	HMM	GMM	ANN	SVM
Average classification accuracy	75.5–78.5%	74.83–81.94%	51.19–52.82% 63–70%	75.45–81.29%
Average training time	Small	Smallest	Back-propagation: large	Large
Sensitivity to model initialization	Sensitive	Sensitive	Sensitive	Insensitive

and [27]

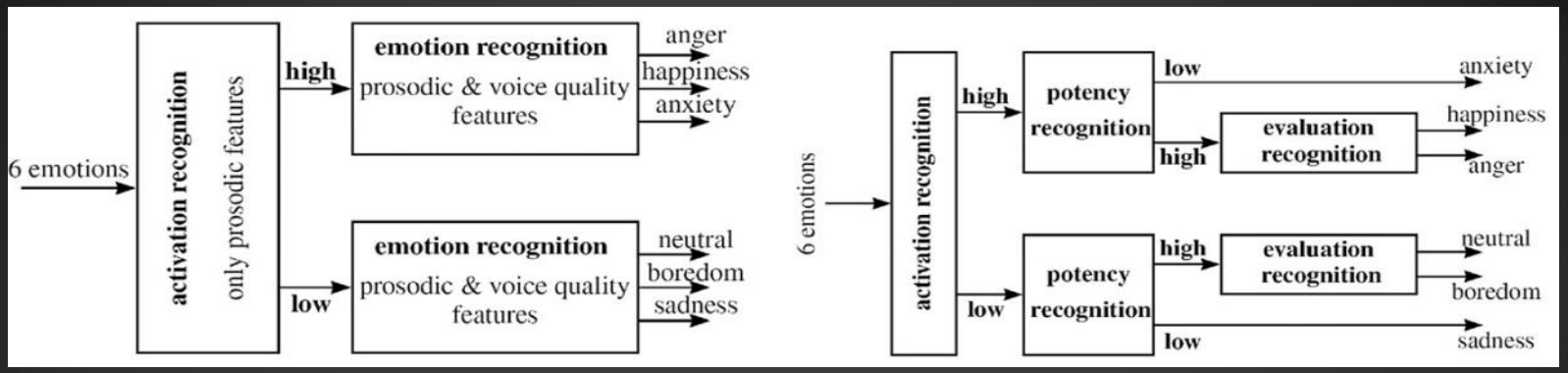
Reference	**Methods Compared**	**Accuracy**	**Speed**	**Computational Efficiency**
Huang et al. (2014) [99]	DNN vs. ELM	DNN: Higher accuracy (5-10% better)	ELM: Faster than DNN	DNN: Higher computational cost
Mirsamadi et al. (2017) [11]	RNN with Attention vs. GMM	RNN: Higher accuracy (up to 10%)	GMM: Faster	RNN: Higher computational complexity
El Ayadi et al. (2011) [219]	SVM vs. DNN, CNN	DNN/CNN: Significantly higher accuracy	DNN/CNN: Slower due to depth	DNN/CNN: Higher resource demands
Zhao et al. (2018) [220]	Deep CNN vs. Traditional Methods	Deep CNNs: Superior accuracy	Traditional methods: Faster	Deep CNNs: Computationally intensive
Trigeorgis et al. (2016) [221]	Multimodal DNN vs. Traditional Models	Multimodal DNN: Significantly better accuracy	Traditional models: Faster	Multimodal DNN: Higher due to multimodal inputs
Poria et al. (2017) [222]	Various Deep Learning Approaches	Deep learning: Advancements in accuracy	Deep learning: Generally slower	Deep learning: Increased data and computational demands

CLASSIFIERS

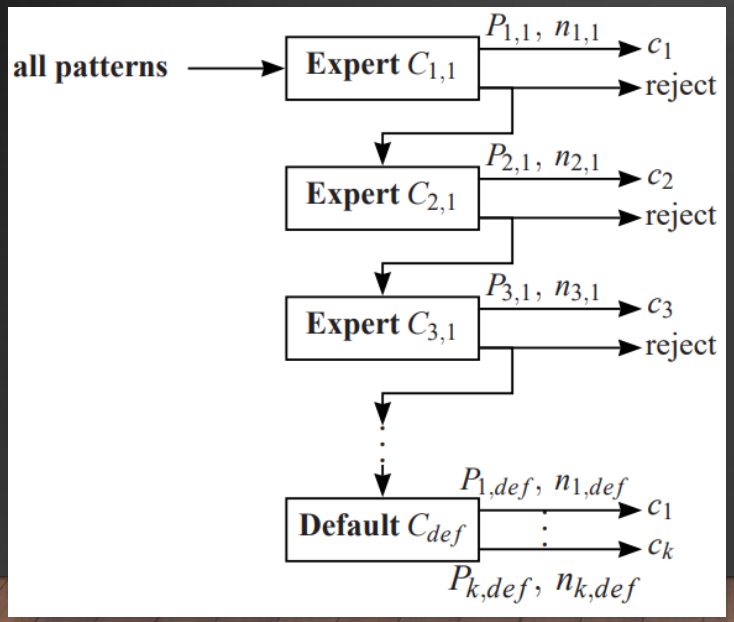
- Combination of classifiers [26]
  - Three approaches
    - Hierarchical: tree structure where candidate classes become smaller as we go deeper in the tree
    - Serial: classifiers in a queue, each classifier reduces the number of candidate classes for the next classifier
    - Parallel: all classifiers work independently and a decision fusion algorithm is applied

CLASSIFIERS

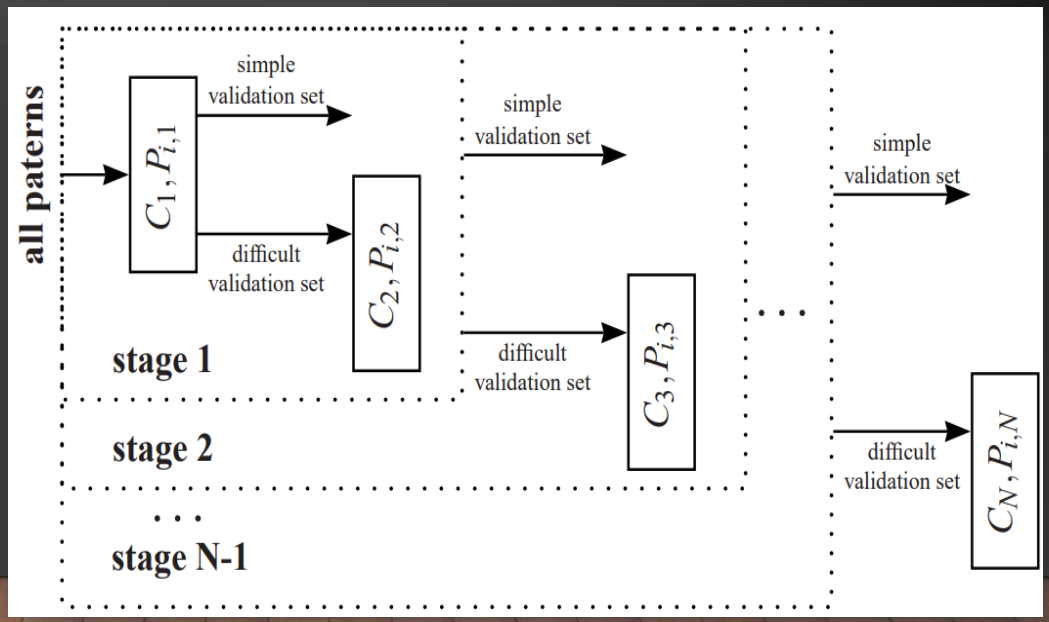
- Hierarchical:



- Serial:



Parallel:



CLASSIFIERS

- DOs and DON'Ts:

✓ **Speaker-independence in train-test splits**

✓ **Speaker-independence in train-validation splits**

✓ **(nested) K-fold Cross-Validation**

✓ **Stratification**

✓ **Ablation study**

X **Report only one or two metrics**

X **Leave statistics aside**

X **Leave uncertainty quantification aside**

X **Use only a single dataset**

X **Ignore state-of-the-art comparison**

OVERVIEW

- Introduction
- Computational Modeling of Emotion
- Acoustic Features
- Classifiers
- **Conclusions**

CONCLUSIONS

- Current body of research focuses on studying many speech features and their relations to the emotional content of the speech utterance.
  - New features are developed
  - Different feature selection strategies
  - No clear conclusion: only one or two databases are tested in each study
- Most of the existing databases are not perfect for evaluating the performance of a speech emotion recognizer.
  - Difficult even for human subjects to determine the emotion
  - Low quality of the recorded utterances → tends to be fixed
  - Small number of available utterances → tends to be fixed
  - Unavailability of phonetic transcriptions → ASR has improved much!

CONCLUSIONS

- N-way classification is still challenging
- No baseline for classifiers, features, datasets
- Emotions in-the-wild
- Domain adaptation/Generalization
- Low resourced languages
- Noise robustness and speaker variability
- Privacy, ethics, demographic bias

QUESTIONS?

REFERENCES

- [1] B. Schuller, G. Rigoll, M. Lang (2004), Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, in: Proceedings of the ICASSP 2004, **1**, 577–580.
- [2] D.J. France, R.G. Shiavi, S. Silverman, M. Silverman, M. Wilkes (2000), Acoustical properties of speech as indicators of depression and suicidal risk, IEEE Trans. Biomedical Eng. 47 (7), 829–837.
- [3] J. Hansen, D. Cairns, Icarus: source generator based real-time recognition of speech in noisy stressful and Lombard effect environments (1995), Speech Commun. 16 (4), 391–422.
- [4] J. Ma, H. Jin, L. Yang, J. Tsai (2006), in: Ubiquitous Intelligence and Computing: Third International Conference, UIC 2006, Wuhan, China, September 3–6, +Proceedings (Lecture Notes in Computer Science), Springer-Verlag, New York, Inc., Secaucus, NJ, USA.
- [5] Scripture, E. (1921). A study of emotions by speech transcription. *Vox*, **31**, 179–183.
- [6] Skinner, E. (1935). A calibrated recording and analysis of the pitch, force, and quality of vocal tones expressing happiness and sadness. *Speech Monographs*, 2, 81–137.
- [7] Fairbanks, G. and Hoaglin, L. (1941). An experimental study of the durational characteristics of the voice during the expression of emotion. *Speech Monographs*, **8**, 85–91.
- [8] Fairbanks, G. and Pronovost, W. (1939). An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monographs*, **6**, 87–104.
- [9] Dellaert, F., Polzin, T., and Waibel, A. (1996). Recognizing emotion in speech. In *Proc. of ICSLP*, pp. 1970–1973, Philadelphia
- [10] Ang, J., Dhillon, R., Shriberg, E., and Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proc. of Interspeech*, pp. 2037–2040, Denver.
- [11] Batliner, A., Fischer, K., Huber, R., Spilker, J., and N¨oth, E. (2000). Desperately seeking emotions: Actors, wizards, and human beings. In *Proc. of the ISCA Workshop on Speech and Emotion*, pp. 195–200, Newcastle, Co. Down.
- [12] Lee, C., Narayanan, S., and Pieraccini, R. (2001). Recognition of negative emotions from the speech signal. In *Proc. of ASRU*, pp. 240–243, Madonna di Campiglio, Italy.

REFERENCES

- [13] H. Schlosberg (1954), "Three dimensions of emotions," Psychological Reveiew, vol. 61, pp. 81-88.
- [14] P. Ekman and W. Friesen (1971), "Constants across cultures in the face and emotion," Journal of Personality and Social Psychology, vol. 17, no. 2, pp. 124-129.
- [15] T. Radek (2011), "How many dimensions does emotional experience have? The theory of multi-dimensional emotional experience," in Re-constructing Emotional Spaces: From Experience to Regulation, Prague, Prague College of Psychosocial Studies Press, 33-40.
- [16] Eyben, F., Wollmer, M., and Schuller, B. (2010). openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. of the 9th ACMInternational Conference on Multimedia,MM*, pp. 1459–1462, Florence.
- [17] Pachet, F. and Roy, P. (2009). Analytical features: A knowledge-based approach to audio feature generation. *EURASIP Journal on Audio, Speech, and Music Processing*.
- [18] Deller, J., Proakis, J., and Hansen, J. (1993). *Discrete-Time Processing of Speech Signals*. Macmillan, New York.
- [19] Furui, S. (1996). *Digital Speech Processing: Synthesis, and Recognition*. Signal Processing and Communications. Marcel Dekker, New York, 2nd edition.
- [20] O'Shaughnessy, D. (1990). *Speech Communication*. Adison-Wesley, Reading, MA, 2nd edition.
- [21] Oppenheim, A. V.,Willsky, A. S., and Hamid, S. (1996). *Signals and Systems*. Prentice Hall, Upper Saddle River, NJ, 2nd edition.
- [22] Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustic, Speech and Signal Processing*, **28**, 357–366.
- [23] Kabal, P. and Ramachandran, R. P. (1986). The computation of line spectral frequencies using Chebyshev polynomials. *IEEE Transactions on Acoustics, Speech, & Signal Processing*, **34**, 1419–1426.
- [24] Hermansky, H. (1990). Perceptual linear predictive (plp) analysis for speech. *Journal of the Acoustical Society of America*, **87**, 1738–1752.

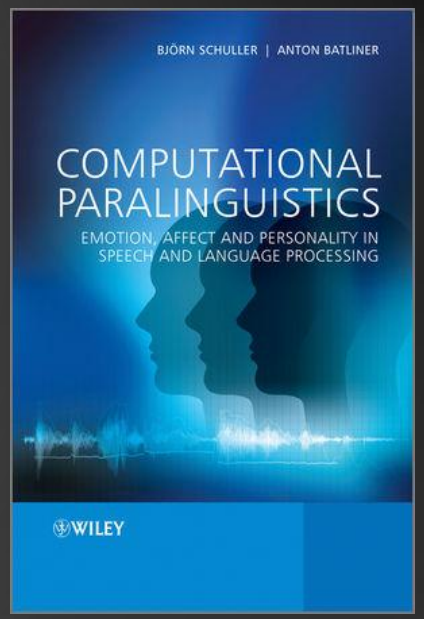
REFERENCES

- [25] Zwicker, E. and Fastl, H. (1999). *Psychoacoustics – Facts and Models*. Springer, Berlin.
- [26] Moataz El Ayadi, Mohamed S. Kamel, Fakhri Karray (2011), Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition, **44**, 3, 572-587.
- [27] G. H. Mohmad Dar and R. Delhibabu, "Speech Databases, Speech Features, and Classifiers in Speech Emotion Recognition: A Review," in IEEE Access, vol. 12, pp. 151122-151152, 2024, doi: 10.1109/ACCESS.2024.3476960.

ACKNOWLEDGMENTS

- Most (if not all) figures in this presentation come from the book

"Computational Paralinguistics, Emotion, Affect, and Personality in Speech and Language Processing", by B. Schuller and A. Batliner, Wiley Press, 2013



- Other books:

1. Analyzing Emotion in Spontaneous Speech, by Chakraborty et al.
2. Real-time automatic emotion recognition from speech: The recognition of emotions from speech in view of real-time applications, by Vogt
3. Emotion Recognition using Speech Features, by Rao and Koolagudi