Exploring Significance of Speech Features for Emotion Recognition

Saurabh Kumar, Akhil Babu Manam, Venkata Aditya Chintala

Motivation

- Emotion recognition is an important problem
- Has applications in many human computer interaction tasks
- Speech contains information related to emotion.

Objective

- Detect emotion of individuals from speech
- Explore acoustic features in speech data
- Classify into categories of:
  - Anger
  - Happiness
  - Sadness
  - Neutrality

Dataset

- Interactive Emotional Dyadic Motion Capture (IEMOCAP)
- Acted, multimodal and multi speaker database
- Collected at SAIL lab at USC.
- 12 hours of audiovisual data, including video, speech, motion capture of face and text transcriptions
- Annotated by evaluators in both categorical and dimensional labels at utterance level
- Categorical labels such as anger, happiness, sadness, neutrality, as well as dimensional labels such as valence, activation and dominance.

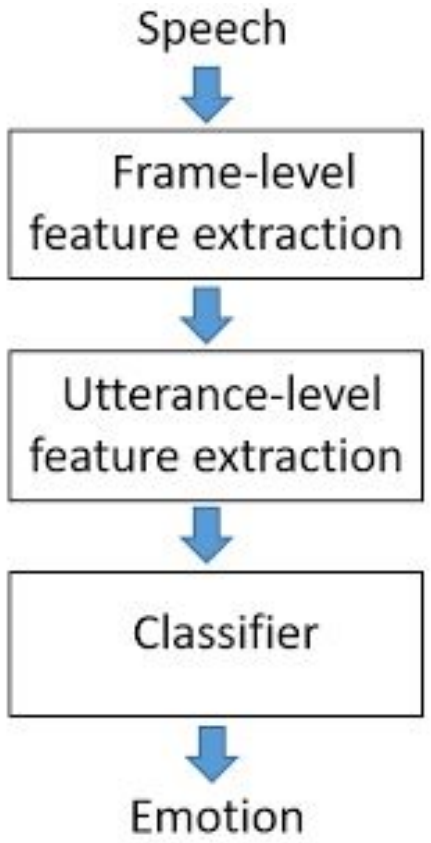
Features

Low - level (Frame - level) features

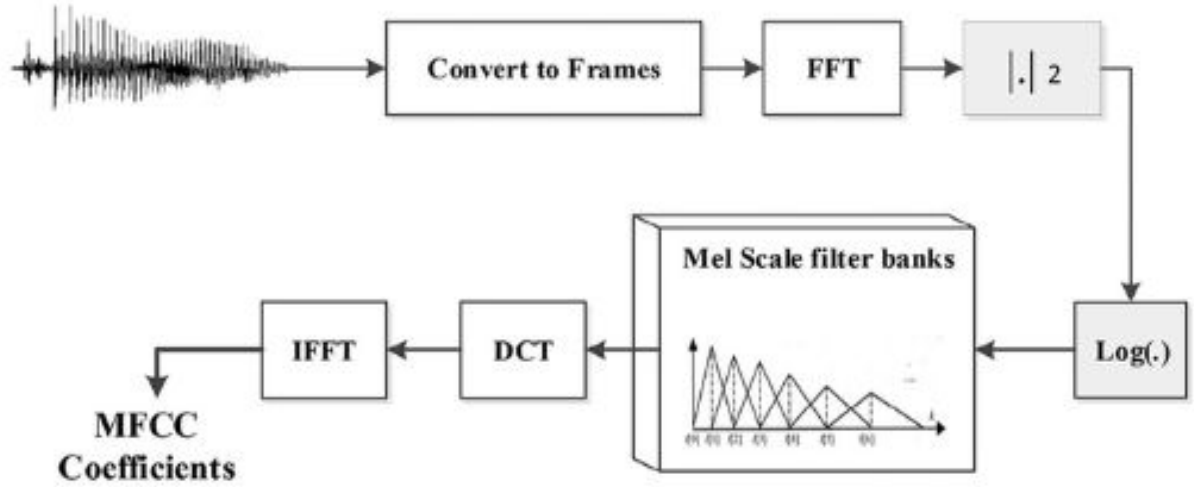
- Mel-frequency cepstral coefficients (MFCC)
- Linear Prediction cepstral coefficients (LPCC)
- Residual Mel-frequency cepstral coefficients (RMFCC)

Utterance - level features

- Shimmer, Jitter
- Harmonics to Noise statistics - (Voice quality)
- Unvoiced to voiced frame ratio
- Opensmile - Paralinguistic Challenge 2010 configuration



Low-level descriptors

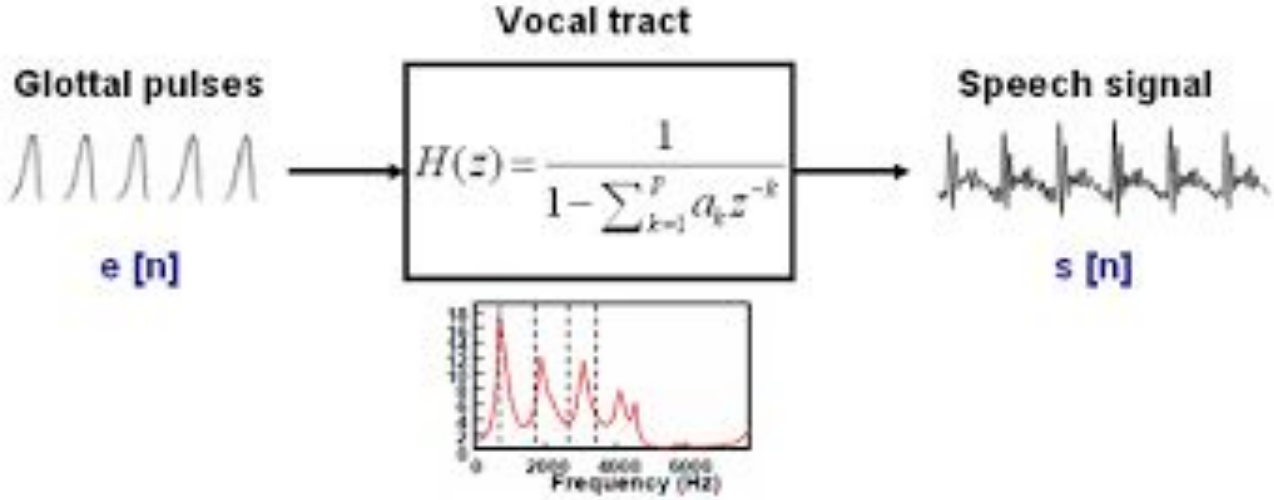


MFCC

- Traditionally used for representing speech for tasks like speech recognition, speaker recognition, emotion recognition etc.
- Contain information related to the vocal tract.

LPCC

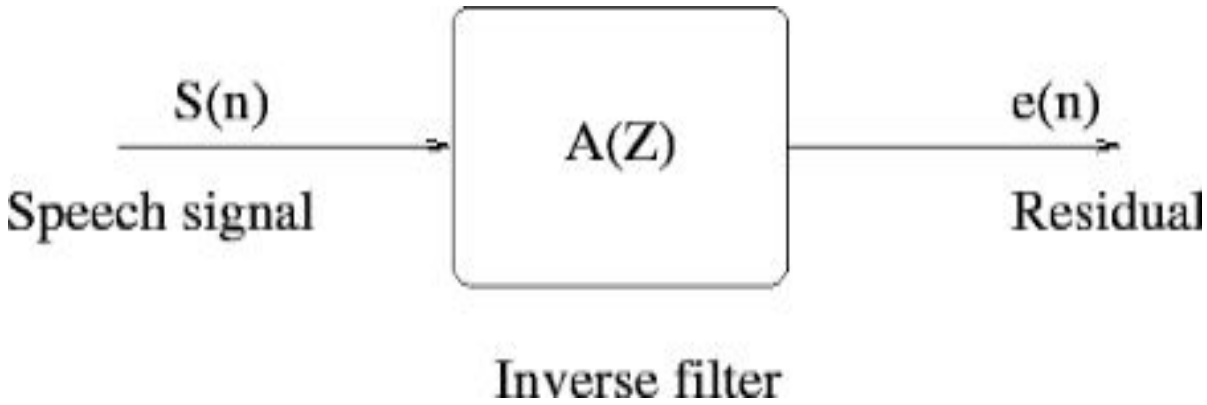
- Also capture information related to vocal tract.
- Used in voice coding.



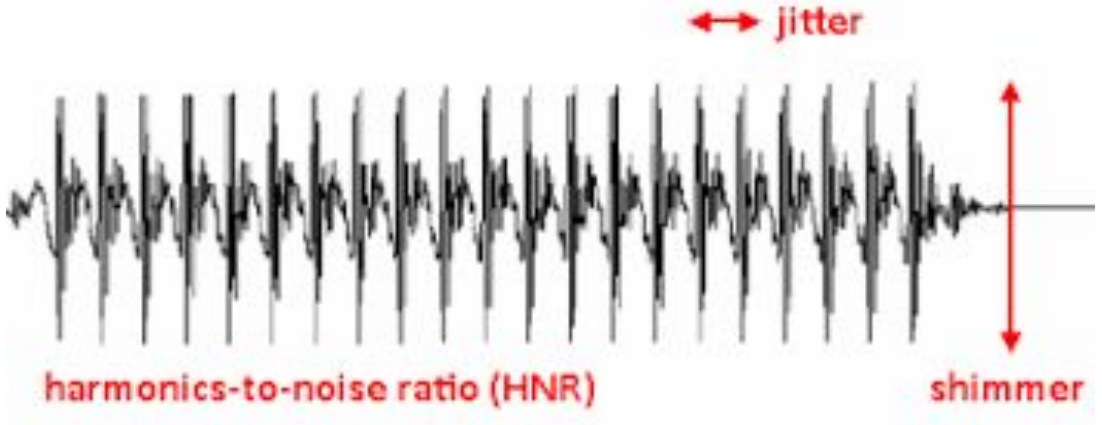
Low-level descriptors

RMFCC

- Residual signal extracted by inverse filtering of signal through LPC filter
- E(z)*H(z) = S(z) => E(z) = S(z)/H(z) = S(z)*A(z)
- S(z) -> Speech signal
- H(z) -> Vocal tract filter (extracted from LPCC coefficients)
- E(z) -> Excitation source approximation



Utterance-level features



Jitter & Shimmer - collected from PRAAT

- Extracted from pitch contour
- Jitter is a measure of sum of $|T(i) - T(i+1)|$, $T(i)$ = time (at pitch i)
- Shimmer is a measure of sum of $|A(i) - A(i+1)|$, $A(i)$ = amplitude

Harmonics to noise ratio - collected from PRAAT

- Harmonicity to noise ratio in the voiced frames

Unvoiced to voiced frames ratio

- Ratio of number of frames that are unvoiced to voiced

Utterance-level features

Opensmile features (OpenSmile Appendix)

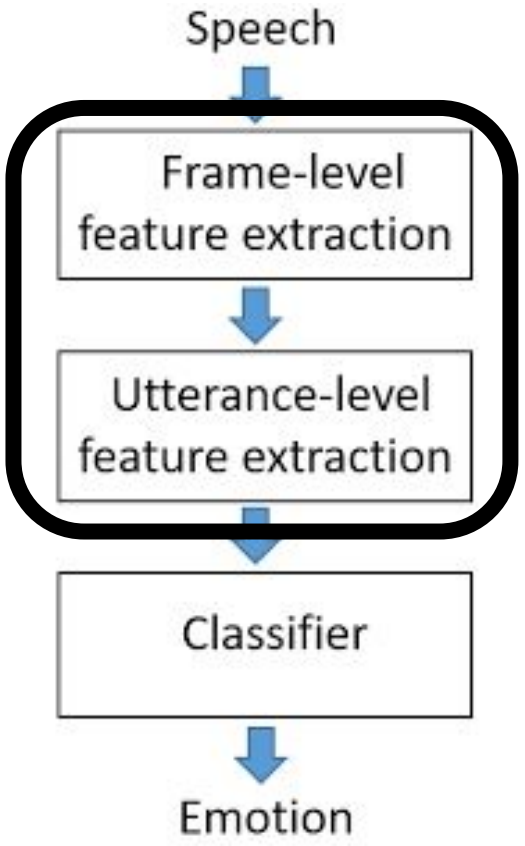
- 1582 features
- Used for Paralinguistic, emotion recognition tasks
- Represent various LLD features using statistics
- Collect various utterance level descriptors

Low-Level Features to Utterance-level

Segment-level representations for low-level descriptors are extracted:

- LSTM categorical embeddings
  - LSTM (512 hidden units)
  - LSTM (256 hidden units)
  - Dense (relu, 256 hidden units) -> feature
  - Dense (softmax, 4 outputs)
- LSTM Autoencoder embeddings

For both these embeddings, the frames are truncated or padded to obtain common shape.



Opensmile vs Deep features

Opensmile features

- Independent of data (pro)
- Cannot be used for complex tasks - Transfer Learning, Multi-task learning, cross lingual emotion recognition etc. (con)

Deep features

- Require enormous labeled data which is expensive and time taking. (con)
- Unsupervised features (LSTM Autoencoders) can be used to solve data sparsity issues (pro).

Evaluation and Results

Leave-one-Subject-out validation

Metrics - Accuracy

Classifiers	**SVM**					
Features	**Weighted (Recall / Precision / F1-score)**			**Unweighted (Recall / Precision / F1-score)**		
Jitter + Shimmer + HNR + UV ratio	31.50 (3.05)			38.30 (3.29)		
LSTM categorical embedding (MFCC)	51.47 (2.20)			**52.75 (4.01)**		
Opensmile features (PC 2010)	**52.70 (1.03)**			48.20 (2.76)		
LSTM categorical embedding	46.73 (11.82)			44.55 (0.16)		

Future Work

- Evaluation on LSTM Autoencoder embeddings
- Check if significant improvement exists by fusing features.

Conclusions

- We have explored various features including deep LSTM features for speech emotion recognition
- Addressed the usefulness of each features in different scenario.
- We have found that their performances are comparable.

References

[1] S.B. Davis, and P. Mermelstein (1980), "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), pp. 357–366.

[2] Atal BS. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. the Journal of the Acoustical Society of America. 1974 Jun;55(6):1304-12.

[3] Ingale AB, Chaudhari DS. Speech emotion recognition. International Journal of Soft Computing and Engineering (IJSCE). 2012 Mar;2(1):235-8.

Questions ?