USING NEUTRAL SPEECH MODELS FOR EMOTIONAL SPEECH ANALYSIS

Carlos Busso, Sungbok Lee and Shrikanth S. Narayanan

Speech Analysis and Interpretation Laboratory (SAIL)
Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089
busso@usc.edu, sungbokl@usc.edu, shri@sipi.usc.edu

USC VITERBI SCHOOL OF ENGINEERING

imsc

Motivation

Why?

- Emotions play a crucial role in human interaction.
- Knowing the users emotional state should help to adjust system performance.
- User can be more engaged and have a more effective interaction with the system.
- Many applications (humanoid robots, call center, educational games, etc).

Emotion recognition in the lab

- In general, limited acted data with few speakers.
- Categorical representation of emotions.
- Many features are selected which are then reduced using feature selection.
- Performance from 50% - 85% depending on the task [1, 2].

Emotion recognition in real life applications

- Too much variability.
  - Speaker variability, acoustic environments, mixture of emotions, etc.
- Models are not easily generalized to other databases or on-line recognition task.

Where is the problem?

- Feature selection.
- Lack of emotional data.
- Over fitting.
- Mismatch between training and testing data.

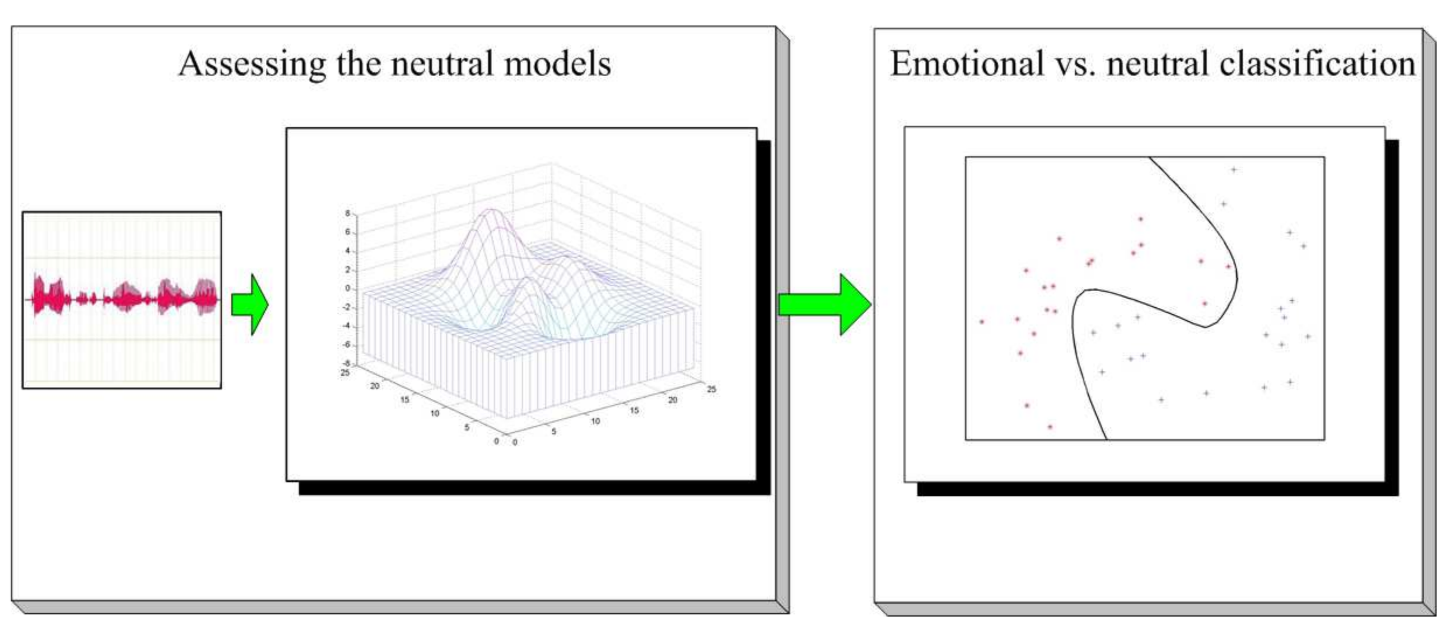
Approach

Hypotheses

- Different emotions are different variants of the neutral emotion.
- Emotion expression affects different speech sounds differently.

Idea

- Discriminate between emotional and neutral speech.
- Acoustic neutral reference models are used for emotion evaluation.
- Build robust models (many emotionally-neutral databases).



Databases

- Reference model are trained with TIMIT corpus (460 speakers, 6300 sentences).
- Two emotional databases.
  - EMA database (Acted, 3 speakers, 800 sentences, neutral, happy, angry and sadness, 16KHz).
  - Call center data (CCD) (Spontaneous, many speakers, 1027 neutral sentences, 338 negative sentences, 8 KHz).

Analysis of the likelihood scores

Spectral acoustic models

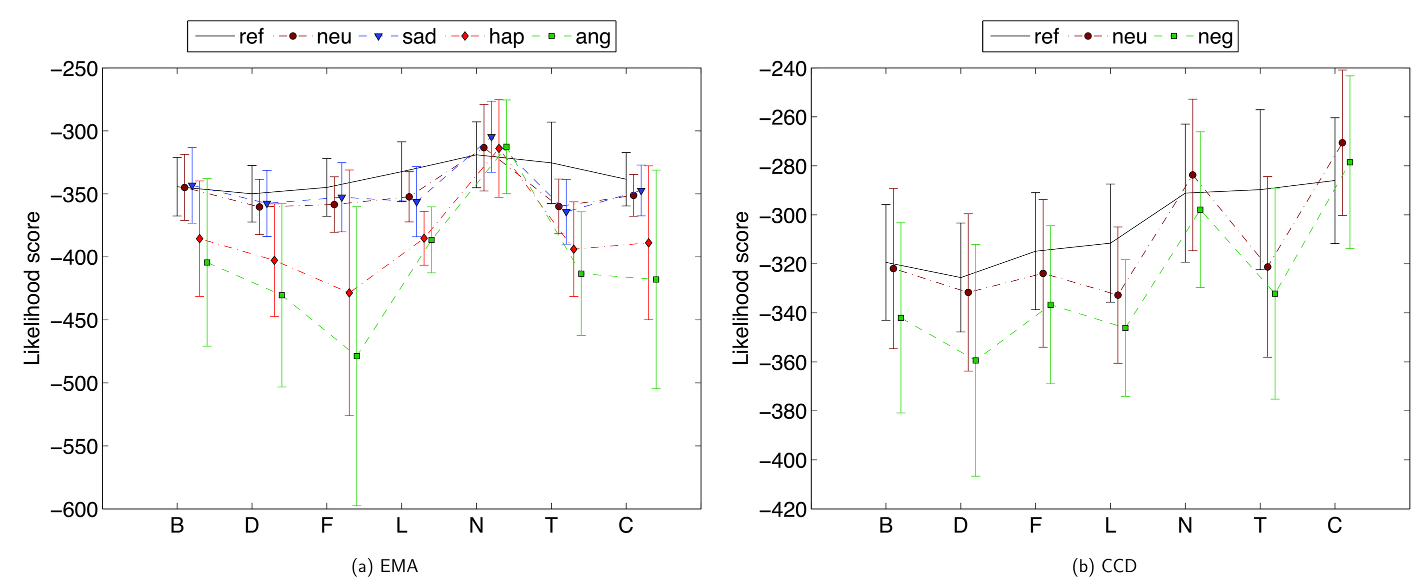
- Spectral features are hypothesized to be advantageous.
- Models are built for MFB and MFCC.

Building models

- HMMs are used to trained broad phonetic classes (3 states, 16 mixtures).
- Normalized likelihood score are used as fitness measurement (Viterbi decoding).
- Energy normalization: E(Neutral set) $\approx$ E(Ref. set).
- For CCD, the TIMIT database was downsampled to 8KHz.

	Description	Phonemes
F	Front vowels	iy ih eh ae ix
B	Mid/back vowels	ax ah axh ax-h uw uh ao aa ux
D	Diphthong	ey ay oy aw ow
L	Liquid and glide	l el r y w er axr
N	Nasal	m n en ng em nx eng
T	Stop	b d dx g p t k pcl tcl kcl qcl bcl dcl gcl q epi
C	Fricatives	ch j jh dh z zh v f th s sh hh hv
S	Silence	sil h# #h pau

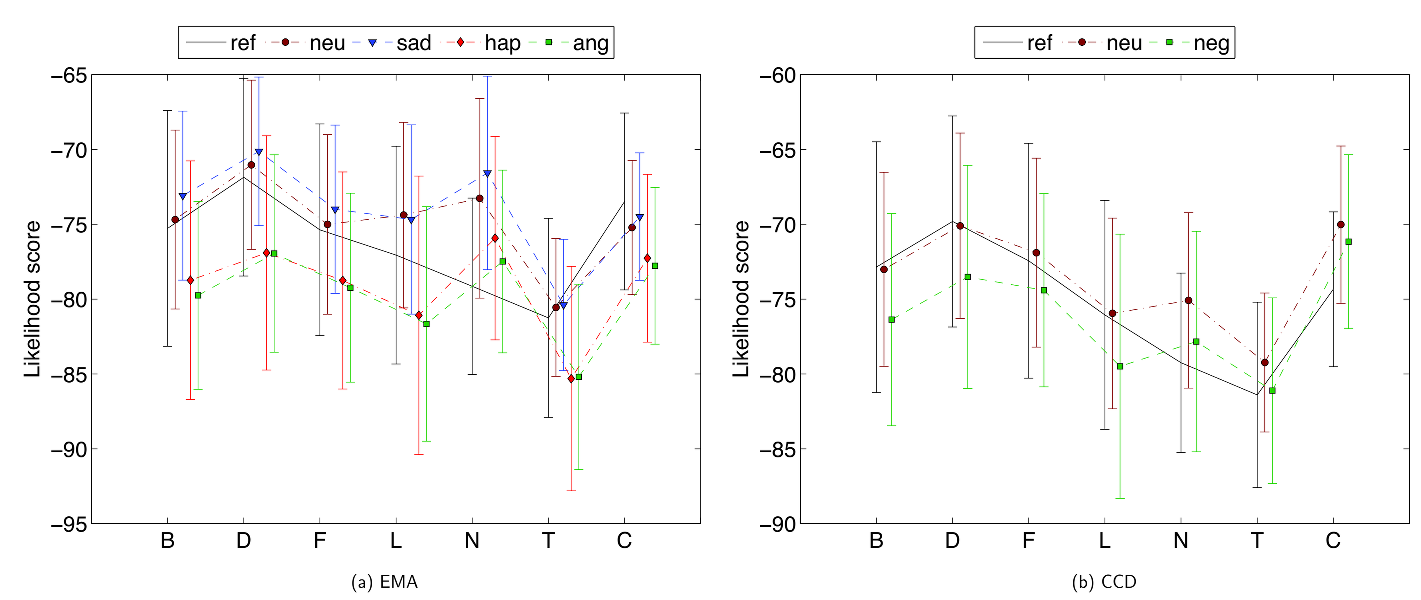
MFB-based neutral speech models



Error bar of the likelihood scores in terms of broad phonetic classes (models trained with MFB).

- The mean and variance of the likelihood score for emotional speech differ from the results observed in neutral speech.
- Strong differences for emotion with high arousal level (i.e., anger and happiness).
- In the valence domain, the differences are not as clear (i.e., sadness vs. neutral).
- Some broad phonetic classes present stronger differences (i.e., front vowels).

MFCC-based neutral speech models



Error bar of the likelihood scores in terms of broad phonetic classes (models trained with MFCC).

- Differences between the likelihood scores for emotional categories are not as clear.
- Emotional discrimination is blured in post-processing steps (i.e, *Discrete Cosine Transform* (DCT)).

Discriminant analysis

- Linear Bayes Normal Classifier (80% training, 20% testing).
- Priors set equal for each classes.
- Features: mean and standard deviation of normalized likelihood scores.
- Results averaged over 100 realizations (product combining results).

EMA

		Ground truth							
		MFB-based models				MFCC-based models			
		Sad	Ang	Hap	Neu	Sad	Ang	Hap	Neu
Classified	Emo	6.4	33.7	33.0	0.6	1.7	33.7	27.9	4.5
	Neu	28.1	0.6	1.5	34.1	33.6	0.5	6.7	29.6
	Sad	21.0	0.0	4.2	9.8	23.4	0.0	0.6	10.6
	Ang	0.1	20.3	14.0	0.5	0.3	21.0	11.9	1.2
	Hap	0.6	7.6	24.3	2.0	1.4	11.4	12.7	9.2
	Neu	9.9	0.0	0.2	23.4	11.8	0.0	1.7	21.0

- 4-label with MFB models EMA: $\sim$ 65% (ref 66.9% [3]).
- Binary classifier
  - MFB models: Acc=0.78, Pre=0.53, Rec=0.98, F=0.69
  - MFCC models: Acc=0.67, Pre=0.42, Rec=0.87, F=0.57
- High accuracy in the activation (arousal) dimension.

CCD

		Ground truth			
		MFB-based models		MFCC-based models	
		Neg	Neu	Neg	Neu
Classified	Neg	38.2	27.4	37.3	25.6
	Neu	52.7	151.7	59.1	144.5

- Recognition rate (ref 74% [4]).
  - MFB models: Acc=0.70, Pre=0.74, Rec=0.85, F=0.79
  - MFCC models: Acc=0.68, Pre=0.71, Rec=0.85, F=0.77
- Most of the sentences are short (noisy mean and std of the likelihood scores).

Conclusion

- Classification performance can achieve accuracy up to 78% (binary classification).
- MFB are better than MFCC.
- This novel approach seems to be suitable for on-line applications.
- Results support the hypotheses.

Future work

- Hierarchical emotion recognition (finer description for emotional speech).
- Reduce mismatch between the reference and emotional corpora.
  - Adaptation of acoustic models.
- Include prosodic features (pitch and energy).
- Use this approach for emotional speech mining in large corpora.

*

References

[1] M. Pantic and L. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, September 2003.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, January 2001.

[3] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. In Press, 2007.

[4] C. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, March 2005.

Acknowledgements

This research was supported in part by funds from the NSF and Army.