Classifying Style Of Spoken Speech

Sunil Kumar Kopparapu, Sathyanarayana Srinivasan, Akhilesh Srivastava, P.V.S. Rao

SunilKumar.Kopparapu@TCS.Com

Cognitive Systems Research Laboratory
Tata Consultancy Services Limited, Navi Mumbai

http://www.tcs.com

November 2006

Speech

- Unique mode of communication (probably) only in the human species
- Speech accounts for nearly 70% of the information and knowledge
- Intelligence in humans is due to language and speech

Information in Speech: Categories

- non-linguistic, (who said it)
  age, gender, anatomy, idiosyncrasy, physical and emotional states

- linguistic (what he said)
  discreet and categorical represented by the written language

- **paralinguistic (how well said*)**
  deliberately added by the speaker, and not inferable from the written text.

*manner, clarity or accent, aspects related to quality

Speech Style Determination: Difficulties

- Qualitative, not necessarily universally uniquely defined.
- Subjective element
  - Self consistency
  - Cross consistency
- Not very easy to capture and characterize.

Self Consistency

Unknown to the expert, 10% of the files were given to him more than once.

Five categories: Very Good, Good, Average, Bad, Very Bad.

Observations

- Exact consistency: 67% (file is given exactly the same tag both times)
- Approx (1-step) consistency: 31% (e.g. same file tagged once as Very Good and at another time as Good)
- Poor consistency: 2% (e.g. once tagged Good and at another time Bad)

Conclusion
Expert is reasonably self-consistent.

This sets the norm for rating system performance

Cross Consistency of Accent Experts

Expert 1 \ Expert 2	VGood	Good	Avg	Bad	VBad
Good	111	461	440	66	28
Avg	127	508	654	104	58
Bad	14	101	268	70	32

Observations

- Poor consistency between experts (because of subjectivity)
- Agree with each other 43.65 % of the time
- Best agreement is for Average category (they agree 45% of the time)

Speech Style Determination: Approaches

- **Approach-1**
  - **human referees categorize them as good, average**
  - **statistical pattern classifier learns (e.g. HMM)**
- Approach-2
  - Choose a few ideal prototypes of excellent speakers,
  - match given speech sample
  - categorize as good, average, etc. depending upon closeness (DTW)

Approach

Speaker to be accessed speaks $\mathcal{W}$ ($|\mathcal{W}|$ =20) predefined words.

Speaking style assessed by *comparing* the test speech sample of the speaker for **every** word $w \in \mathcal{W}$ with the corresponding statistical models (HMM) of w from $\mathcal{G}$ groups (Good, Avg, etc)

The comparison would give a measure of the closeness of the test sample to each of the $|\mathcal{G}| = 5$ categories.

The closest (group) score for all the $w \in \mathcal{W}$ to each of the classification groups is combined to come out with an overall category classification.

Need to build statistical models

Building Statistical Models

Training samples collected from several people (in different acoustic conditions) and classified by a human expert into one of the $\mathcal{G}$ categories.

Statistical model of each word w for each classification category $\mathcal{G}$ is constructed using training samples.

Parameters, (both articulatory ($\mathbb{D}$) and intonation ($\mathbb{I}$)), are extracted and used to build HMMs.

$\mathbb{D}$ consists of 12 MFCC (Mel Frequency Cepstral Coefficient), 8 Δ MFCC, 4 Δ^2 MFCC, while $\mathbb{I}$ consists of variation in pitch and variation in amplitude.

Experimental Speech Data

- 60 speakers
- 20 words, 10 sentences, 3 repetitions, (= 5400 speech files)
- Classified by an accent expert
- Classification Categories
  - Very Good, Good, Average, Bad, Very Bad

Cross Consistency: System Vs Expert

Performance (file level classification)

- Train Data
  - 94% (Male)
  - 94% (Female)

- Test Data
  - 64% (Male)
  - 65% (Female)

Much better than agreement between two human experts (43%)

Experiment with Real Life Data

- 30 candidates
- 5 level tagging (Very good, Good, Average, Bad, Very bad)
- Candidates rated independently by a human expert and the (pre-trained) system.

The human did not know the machine rating during the course of the session.

We compared the performance of the human expert against the system.

Evaluation Method: Candidate Rating

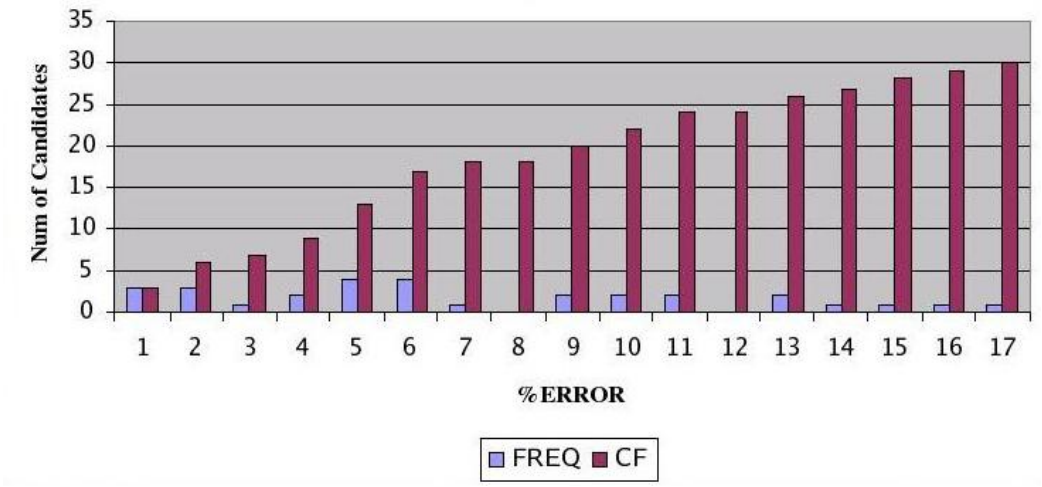
The scoring for each utterance is done as follows:

5 $\rightarrow$ Very Good; 4 $\rightarrow$ Good;
3 $\rightarrow$ Average;
2 $\rightarrow$ Bad; 1 $\rightarrow$ Very Bad

A composite score (range 0 - 100) is obtained based on the scores for each of the 20 utterances adding individual scores

Person Category	**Overall Score**
Very Good	81 - 100
Good	61 - 80
Average	41 - 60
Bad	21 - 40
Very Bad	0 - 20

Result Summary: Candidate Rating



For 24 - out of 30 speakers (80 %) the difference in scores between the human expert and system falls within 10 points; **only half a bin width for 80% of the speakers.**

Score differences are within 1 bin width for **all** speakers. **High agreement** between the human expert and our system.

Conclusions

- Problems here similar to those in speech and speaker recognition.

- Progress is to level of real life application (subjectivity is removed)

- Our metric captures
  – articulatory (source) as well as
  – intonation (system) related aspects

- **Content independent** quality evaluation tempting goal appears to be reachable, but not easy.