



# Trends in Signal Processing

# Speech and Language Processing



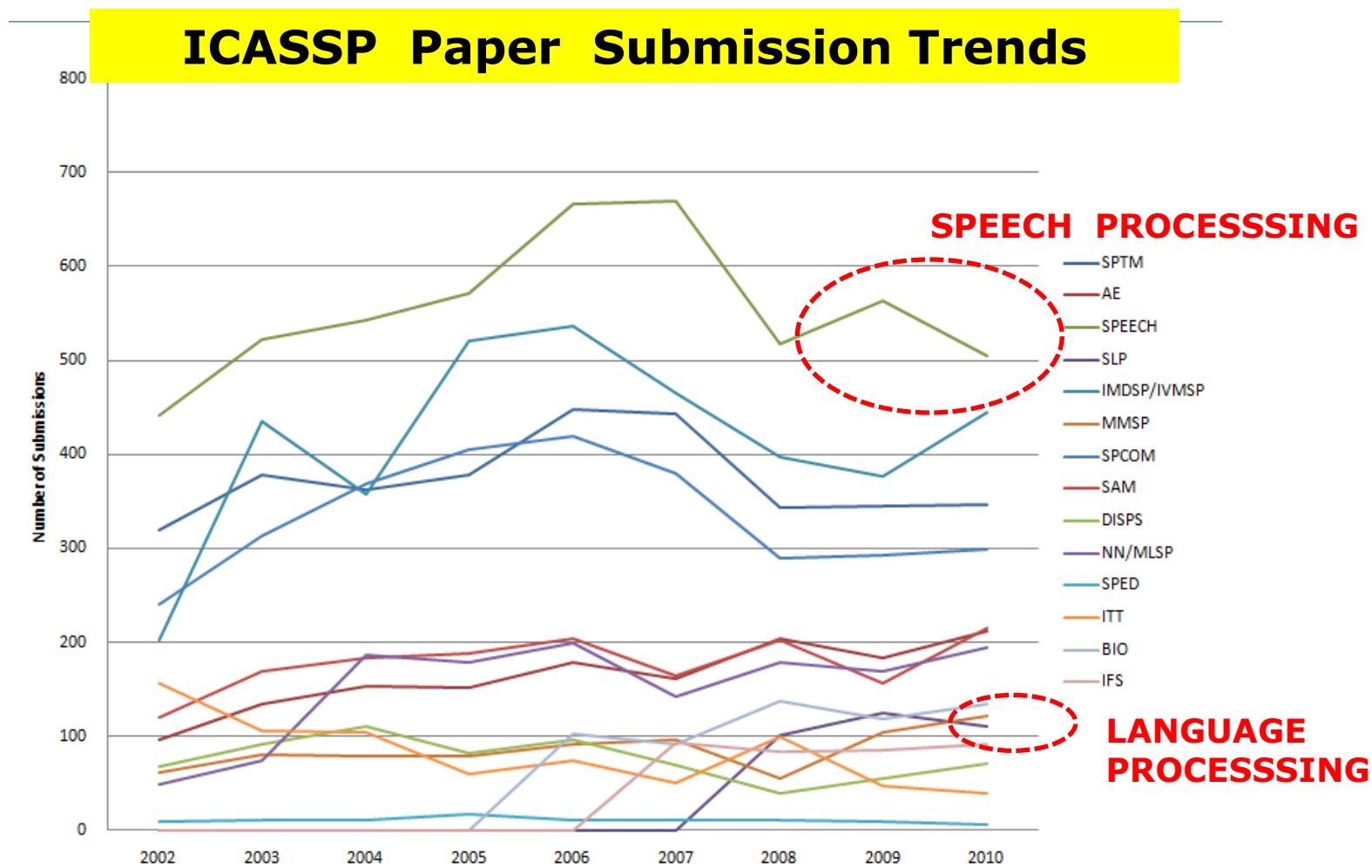
**John H.L. Hansen**

**Friday, May 27, 2011**

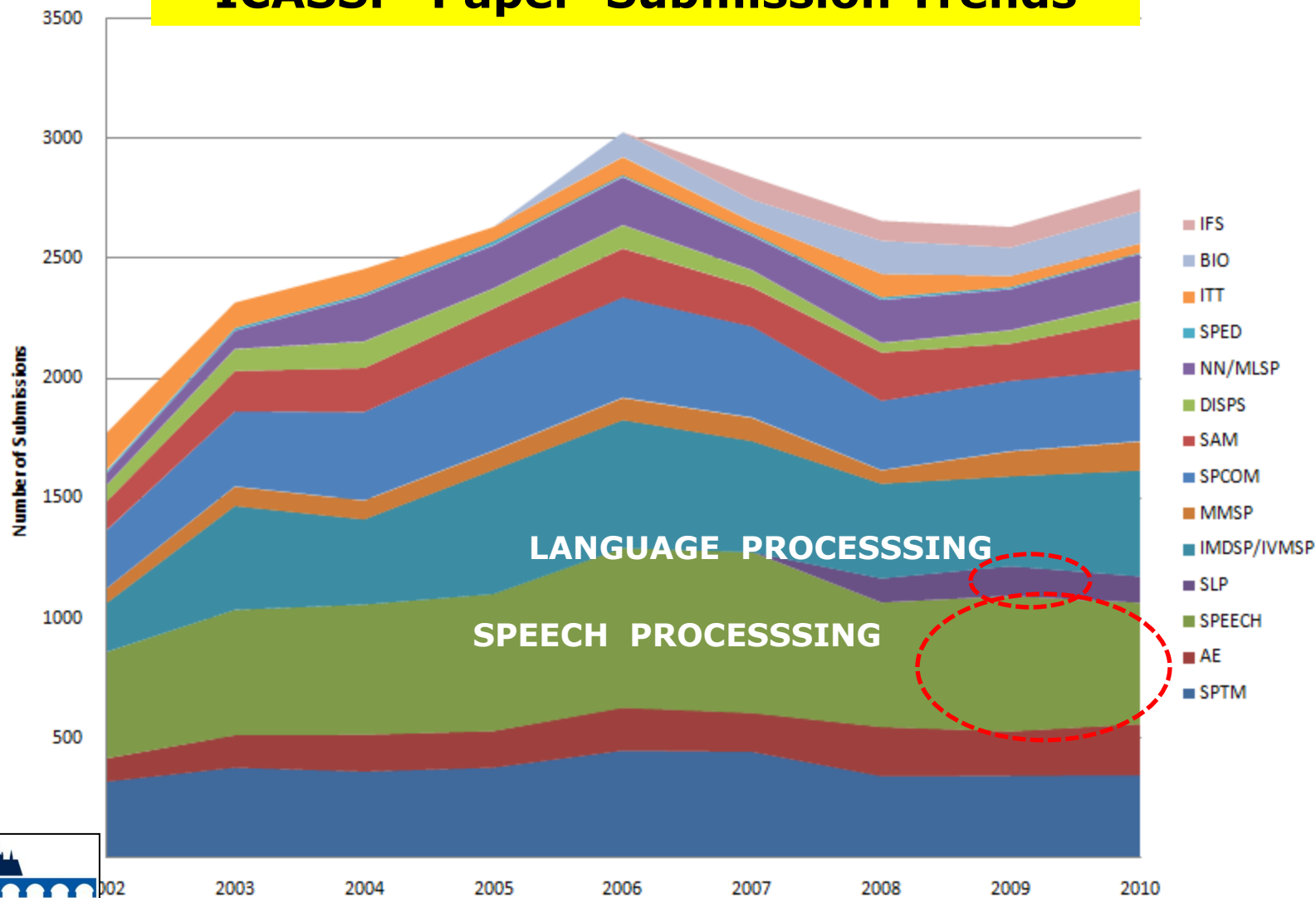


## Outline

- ◆ **General Trends in Speech & Language Papers at ICASSP: still the largest area(s)**
- ◆ **Introduction of Area Speakers for today's Session**
- ◆ **Overview Article to appear in IEEE Signal Processing Magazine from today's presentations and discussion**



## ICASSP Paper Submission Trends



## Agenda:

- ◆ IEEE SPS Trends in Signal Processing
- ◆ Speech & Language Processing (Friday, 11:45-12:15)



**Junlan Feng,  
AT&T Labs-  
Research**



**Bhuvana Ramabhadran,  
IBM T.J. Watson  
Research**



**Jason Williams,  
AT&T Labs-  
Research**

**Overviews provided by Junlan, Bhuvana, and Jason!**

**We encourage questions & discussion!**



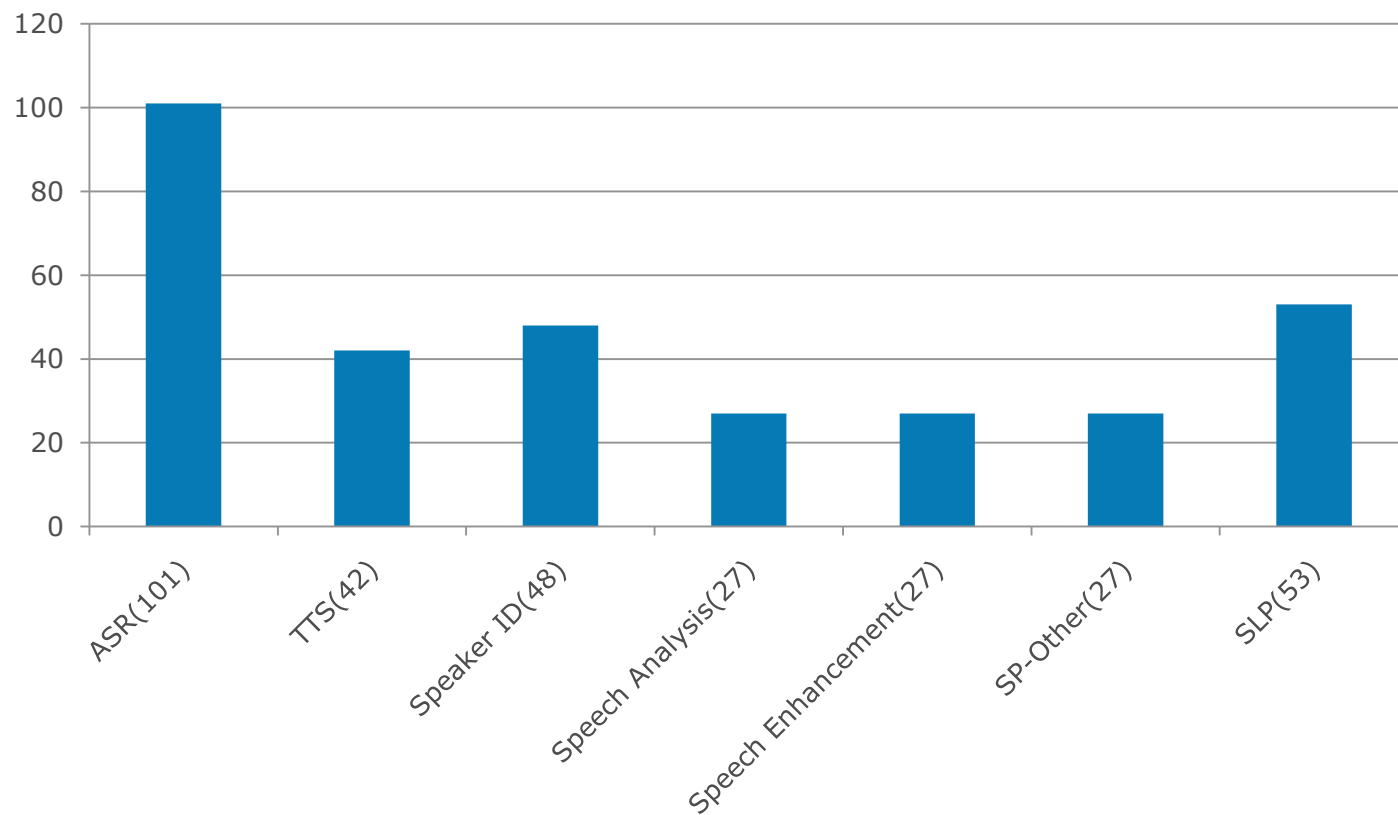
# Trends in Speech and Spoken Language Processing

*Junlan Feng  
Bhuvana Ramabhadran  
Jason William*

# Acknowledgement

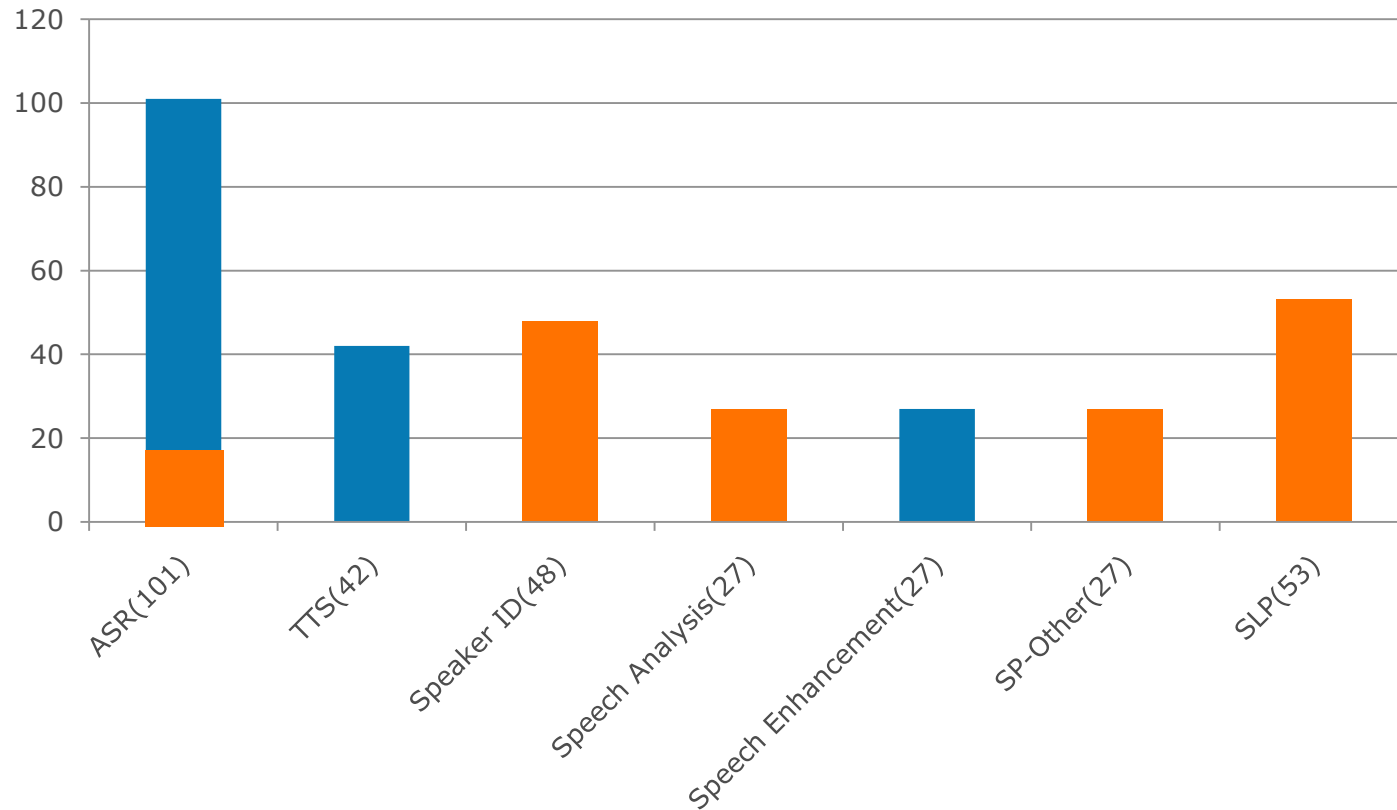
John H. Hansen, Frank Soong, Peter Li  
All SLTC committee members

# 325 Papers



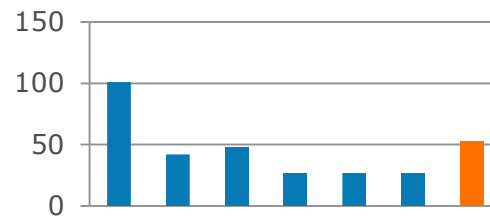
**83% on Speech; 17% on SLP**

# 325 Papers



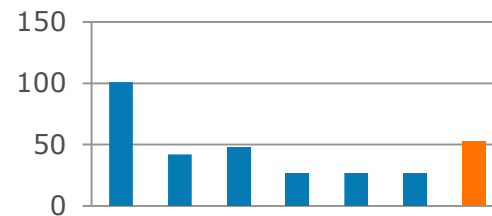
**83% on Speech; 17% on SLP**

# SLP: Language Modeling (11 papers)



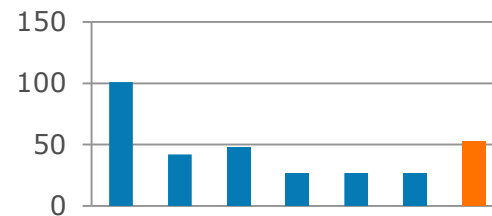
- Approaches for Higher Performances
  - Exponential LM: Model M, a class-based exponential LM
  - Neural Network LM (NNLM): Structured Output Layer Neural Network LM, Recurrent NNLM training
  - Long-Span models
  - Dynamic LM adaptation: Relevance LM
  - Discriminative LM: Round-Robin
- Computing Optimization:
  - Distributed Training
  - Faster NNLM training
  - Manageable Long-span LM
- Data Sets:
  - Gale , NIST, WSJ Data Sets

# SLP: Spoken Document Processing (6 papers)



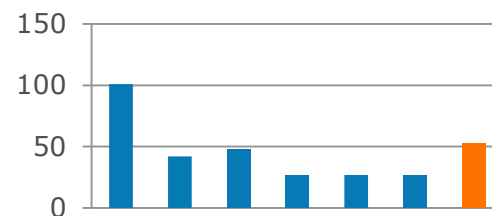
- Tasks of Interest:
  - Document Summarization, Classification, Speaker Role Identification
- Approaches
  - Applying LDA, CRF and other Machine Learning classifications

# SLP: Translation and Semantic Classification (12 papers)



- Tasks of Interest:
  - Query (spoken language) understanding in voice search
  - Language understanding:
    - Deep-Neural networks (DBNs) for call routing
  - Speech Translation
  - Bilingual audio-subtitle extraction

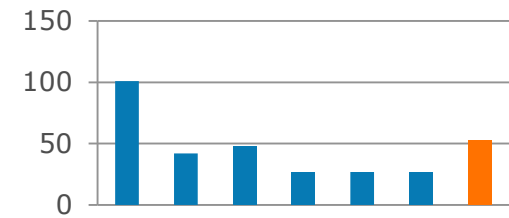
# SLP: Paralinguistic and Non-Linguistic Features (10 papers)



- Tasks of Interest:

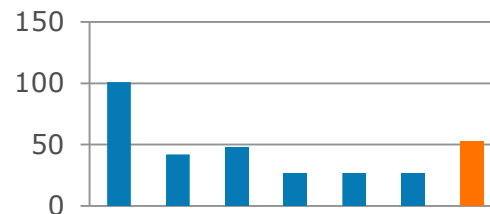
- Emotion detection,
- Recognizing non-lexical Yes or No,
- Cognitive load classification,
- Perceptual difference of phonemes between native and nonnative speakers
- Visual speech synthesis,
- Generating avatar's facial expressions

# SLP: Spoken term recognition and language understanding (10 papers)



- Tasks of Interest:
  - Spoken Term Detection: (return a list of spoken utterances containing the term requested by the user)
  - Mispronunciation Detection
- Approaches
  - Sub-word recognition and retrieval, Dynamic Time Warping, graph-based approaches
- Data:
  - NIST-STD Data sets

# SLP: Dialog (5 papers)



**Recent trend:** Statistical approaches

**Idea 1:** Track a distribution over all possible dialog states

- Allows all information on the N-Best list to be used
- Whole-dialog confidence scores
- Principled use of a prior distribution on user goals

**Idea 2:** Choose system actions using reinforcement learning (RL)

- Allows much more detailed dialog plans to be created
- Allow designer to focus on setting high-level system goals

When these two ideas are used together, the result is a *Partially Observable Markov Decision Process (POMDP)*

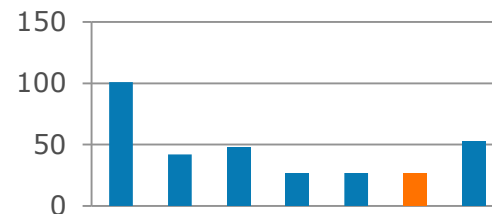
**Recent progress/current challenges:**

- Several techniques now exist to scale to real systems; real systems have been built and shown to yield improvements
- Focus now is on finding good features for RL; methods for training models and optimizing RL; good RL cost functions; how to integrate taxonomies and world knowledge into statistical frameworks; dialog simulation

**Also noteworthy:** Turn-taking; affect/emotion/interest detection; language generation

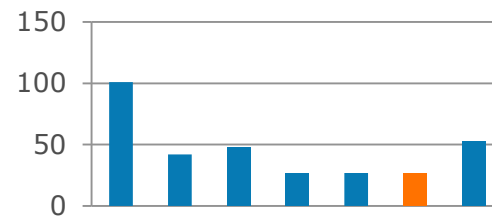
**Relevant session:** SLP-P1: Dialog Systems and Language Modeling I

# SP - Other: Language Identification (6 papers)



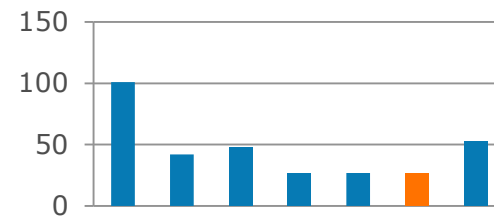
- Features:
  - Phonetic features, prosodic features, a combination of both
  - Low-level audio and visual features in music videos
- Approaches:
  - Classification such as logistic regress, N-grams
- Data:
  - NIST LRE 2009, 2007
  - Singing data in various languages

# SP – Other: Lexical Modeling (10 papers)



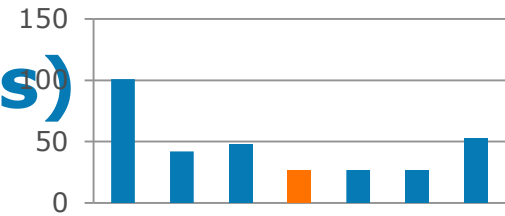
- Approaches:
  - CRF models for grapheme to phoneme (g2p),
  - Machine-Translation based approaches (g2p)

# SP—Other: Multi-Lingual and Multi-channel processing (11 papers)



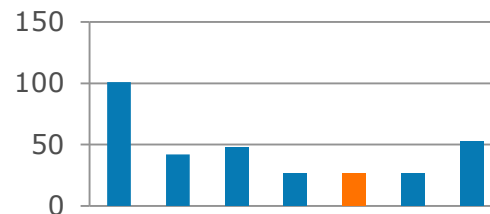
- Tasks of Interest:
  - Mixed Language ASR,
  - Multilingual index and search,
  - Multiple microphones in meeting recognition,
- Approaches
  - Very diversified

# SP: Speech Analysis(27 papers)



- Tasks/Problems of interest:
  - Emotion Detection( sentence level, F0 range vs. Emotions, Emotion vs. SV/SID, Anger detection),
  - Duration Modeling for LVCASR, Pitch Frequency Estimation, Impact of varying types of noises on ASR, Phonetic segmentation , Quantifying Perturbations, Hearing loss simulation, TURBULENCE-NOISE COMPONENT ESTIMATION
- Approaches:
  - Singularity exponents (SE), CRF, phase locked loops(PLLs)
- Data
  - TIMIT,
  - 'Universal Access' database of spastic dysarthric speech

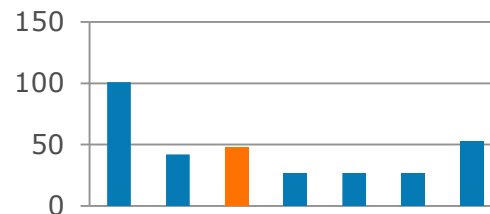
# SP: Speech Enhancement (27 papers)



- Tasks of Interest:
  - (co-channel)Speech separation, noise power spectral estimation, phoneme selective speech enhancement, noise correlation matrix estimation, music noise reduction
- Approaches & Features:

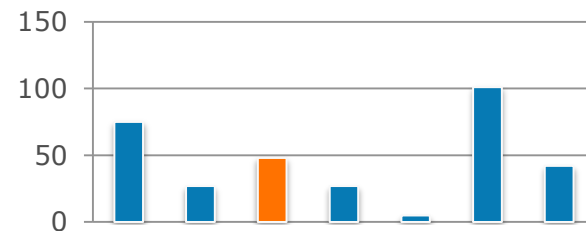
Advances on Wiener Filtering, Improved Kalman filtering, GFCC (Gammatone frequency cepstral coefficients), Spectral Envelope Model,, binary mask Minimum Mean Square Error (MMSE) estimators, Data-Driven Residual Gain Approach Multi-band spectral subtraction, phase modification, maximum a posteriori estimation jointly of spectral amplitude and phase (JMAP).

# SP: Speaker Verification /Identification (42)



- Feature Space , Signal Processing
  - SV/SI in **I-Vector Space**, Prosodic features, Hibert Envelope-based features, CASA Front-End, GIBBS sampling, channel compensation approaches such as Joint Factor Analysis (JFA), Parallel transformation network features
- Models , Approaches:
  - Advances for vocal track model, **PLDA**, Classification, Bayesian, Discriminatively training, GMM, Partial Least Squares , structural Map adaptation, clustering , **System fusion**
- Applications
  - Detection of synthetics speech, Channel-blind SV system, SV/I for gaming,
- Data & Evaluation:
  - MIT LL 2010, NIST Evaluation Data Set for SV, TIMIT, NIST SRE

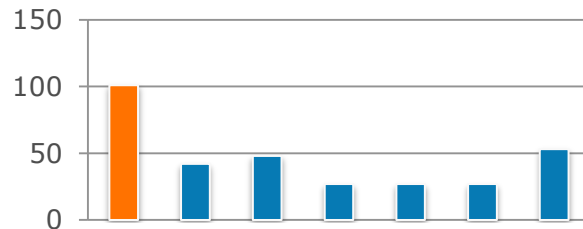
# SP : Speaker Diarization



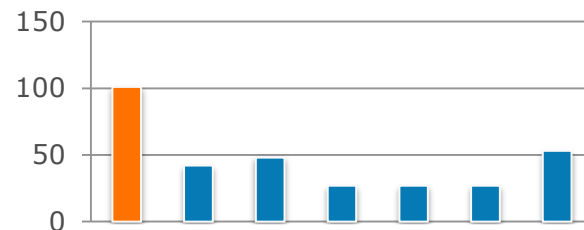
- The task:
  - Infers *who spoke when* in an audio stream or a meeting
- Application:
  - speaker diarization in meetings, web videos
- Approaches:
  - Top-down and Bottom-up clustering, Beyond acoustic features: Speaker role N-gram models, binary keys, Information Bottleneck-based approach

# SP: Robust ASR (28 papers)

- Signal Processing:
  - Compressive sensing
  - NON-NEGATIVE MATRIX factorization (NMF)
  - Discrete wavelet transform (DWT) based filter banks,
  - Others: Factor analysis, Noise Estimation, Cross-channel spectral subtraction, noise estimation, iterative least-square techniques for dereverberation , non-linear noise compensation using gauss-newton method, MULTI-MICROPHONE INTERFERENCE SUPPRESSION
- Features for robust ASR:
  - Features based on Tendem (discriminately trained using multi-layer-percetron or Deep Belief networks),
  - Feature sequence mapping
  - Noise feature normalization
- Models:
  - Dynamic noise adaptation(DNA), MLLR + VTS(vector Taylor series)
  - Others: ML Adaptation of Histogram Equalization, Frame-wise HMM Adaptation, Sampling-based environment population projection for rapid AM adaptation, structured discriminative models, multiple prior models, multimodal training



# SP: Adaptation for ASR (6)



**Problem:** How do you adapt models or feature space to a speaker and the environment with little or moderate amounts of data?

**Recent trend:** Enforcing sparsity or structure on the transforms learnt, improved optimization techniques

## **Ideas:**

Discriminative feature space transforms, Defining parameters as a function of the amount of test speaker data, Prevent overtraining in traditional methods by introducing constraints on the subspace, Rapid adaptation techniques that quickly achieve good performance with few seconds of data as if trained on much more data, Use of basis vectors learnt from the training data

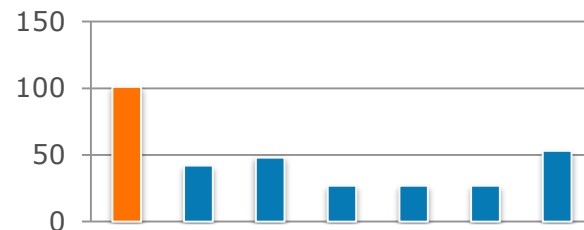
## **Recent progress/current challenges:**

Very good results with single sentence (two or three words) in tasks such as open voice search

Focus now rapid adaptation for real world tasks, incorporation of convex optimization methods, basis representations

**Relevant session:** SP-L3: Adaptation for ASR

# SP: Modeling for ASR (10)



## **Problem: Statistical Modeling for better ASR**

**Recent trend:** Machine Learning techniques, Speeding up learning algorithms to scale to large quantities of data; More and more focus on real world applications (such as voice search)

## **Ideas:**

Capture long-range context via recurrent neural networks

Use of posteriors in multiple stream combination or as features

Return to phoneme recognition with the goal of investigating new techniques rapidly

Posteriors derived from classifiers modeling certain time-frequency bins (learn structure automatically as in RBMs)

Acoustic units (back to syllables?) Range depending on the language? Better units now that amount of data is huge?

Use of language ID/accent and dialect identification/acoustic and phonotactic features for modeling ASR; non-contextual features in acoustic decision trees

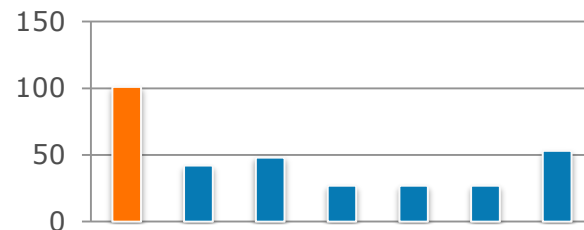
## **Recent progress/current challenges:**

Application specific Loss function for techniques such as gradient boosting

Subspace GMMs for compacting acoustic models

**Relevant session:** SP-P6: Modeling for ASR

# SP: Modeling for ASR (20)



**Problem: Statistical Modeling for better ASR**

## **Ideas:**

Phoneme posteriors as event detectors, Modeling posteriors (HMMs or NNs ) using different distributions; Segmental CRFs in the JHU workshop to fuse multiple sources of information (template based, posteriors based)

Complex models such as long -span LMs and AMs require efficient lattice rescoring/decoding algorithms, CRFs, HMM combinations, multiple streams

More of Neural Nets- bottleneck features, tandem approaches, Deep Belief Nets

Speaker recognition using forensic approaches, Extension of optimization methods to objective functions in new paradigms, Decoding strategies

## **Recent progress/current challenges:**

State dependent basis vectors (BS-HMM), Sparse representations, Exemplar based methods

Capture High order statistical structure via RBMs and DBNs

Use of NMF for unsupervised vocabulary acquisition using acoustic cooccurrences

Discriminative Pont process models for capturing spectro temporal patterns

**Also noteworthy: Relevant session:** SP-P10: Statistical Methods for ASR, SP-P13: Miscellaneous ASR

# SP: Modeling for ASR (10)

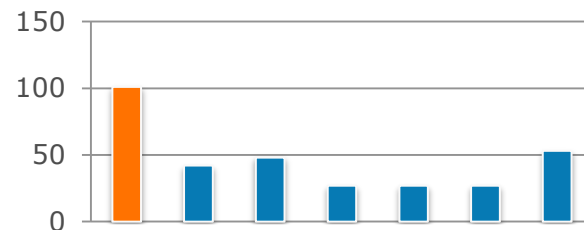
**Problem: Statistical Modeling for better ASR  
(focus on discriminative training methods)**

**Ideas:**

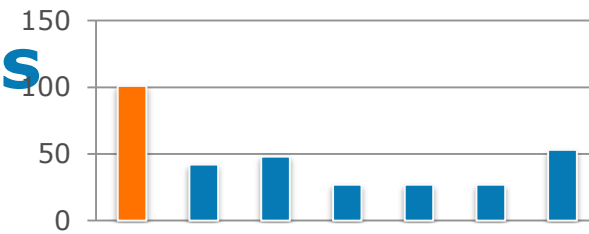
Discriminative training for Full covariance models, MAP, and BSHMMs,  
Feature selection vs l1 regularization of log linear models for speed ups in  
training,

Training criteria that directly minimize WER , Regularization on MLP, parametric  
modeling of durations

**Relevant session: SP-P17: Discriminative techniques for ASR**



# SP: Large Vocabulary Continuous Speech Recognition (10)



**Problem: Statistical Modeling for better ASR**

**(focus on large vocabulary ASR systems)**

**Ideas:**

Unsupervised training (lattice-based)

Recipes Acoustic and Language Models for Mandarin and Arabic GALE evaluations (better context modeling, Morphological features in Neural Network based LMs, Neural network features, Model M, system combination)

Unit selection for LVCSR (Korean, Polish)

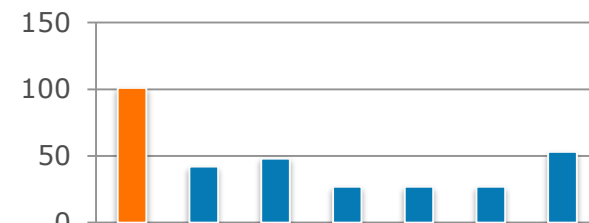
Improving transcripts for unsupervised learning

Decoding schemes that exploit structure of LMs

Exploration of Deep Belief Nets for LVCSR voice search tasks

**Relevant session: SP-P1: Large Vocabulary Continuous Speech Recognition**

# SP: Acoustic Modeling(15)



**Problem: Front-end : Alternative features for HMMs, Modeling paradigms**  
(focus on phone recognition and large vocabulary ASR systems)

## Ideas:

Taeger-Kaiser energy based features, Articulatory trajectories, Pitch-adapted non-stationary features, Tone and Pitch detection, deal binary masks

Efficient parameter estimation by capturing phonetic variability in Sub-space GMMs (JHU workshop 2009).

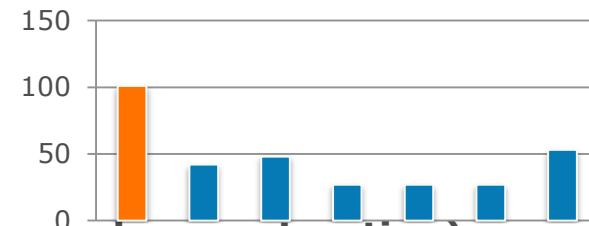
Log-linear models, Full-covariance models, Multi-stream Gabor filtering based features

Kernels/Methods to deal with non-convex optimization issues in neural networks for ASR

Overlapped speech detection and non-audible murmur detection

**Relevant session:** SP-L5: Acoustic Modeling I; SP-P14; Acoustic Modeling II

# SP: Speech Synthesis (42)



**Problem:** HMM based synthesis (underlying parameterization and reconstruction),  
Machine learning Techniques for prosody, pitch accent and boundary detection,  
Concatenative TTS

## Ideas:

Excitation Modeling for HMM based TTS, Voice Conversion, Rapid Voice Adaptation

Embedded HMM based TTS systems (fixed-point arithmetic)

Emotional speech generation

Impact of machine translation (fluency) on speech synthesis

Parameter Tying

Modeling prosody in concatenative TTS systems, Tagging (rules and data-driven methods) to improve TTS, Active learning for pitch accent and boundary prediction, Improved F0 modeling

Constraints for improved unit selection in concatenative TTS systems

**Relevant session:** SP-P2, P12, L8, P18: Speech Synthesis I- IV, MLSP-P5: Machine Learning for Speech and Audio Applications

# IEEE 2011 WORKSHOP ON AUTOMATIC SPEECH RECOGNITION AND UNDERSTANDING

Hilton Waikoloa Village, Big Island, Hawaii

11-15 December 2011

[www.asru2011.org](http://www.asru2011.org)

**Paper submission deadline**

Paper acceptance/rejection

Early registration deadline

Workshop

**1 July 2011**

20 August 2011

15 October 2011

11-15 December 2011

 **IEEE**  
Advancing Technology  
for Humanity

 **IEEE**  
Signal Processing Society