

# Rapid Speaker Normalization with Limited Data using the Second Subglottal Resonance

Shizhen Wang

Advisor: Prof. Abeer Alwan

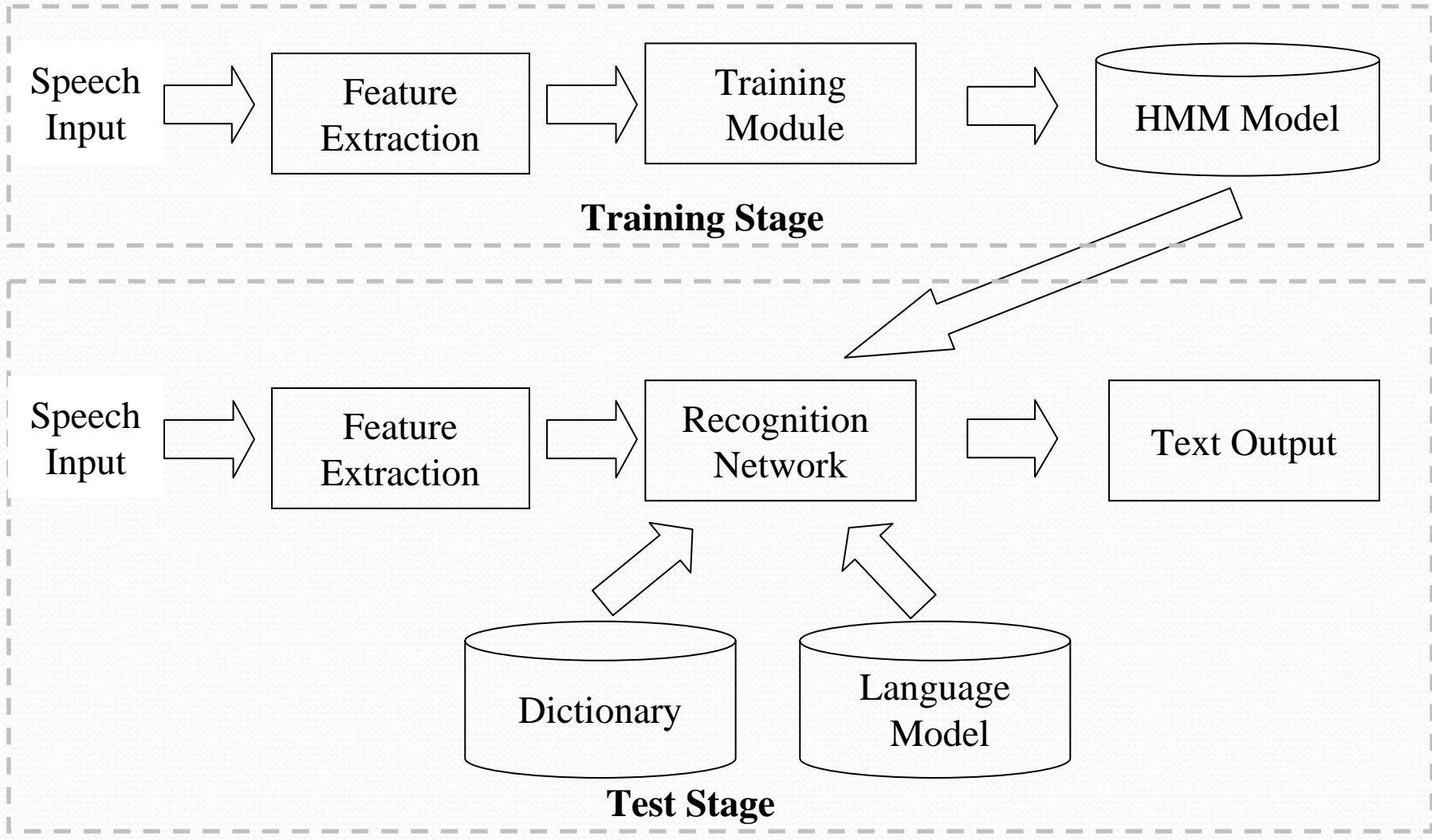
# Outline

- Introduction
- Subglottal system and resonances (Sg)
- Automatic detection of Sg's
- Experimental results
- Conclusions

# Introduction: Automatic Speech Recognition (ASR)

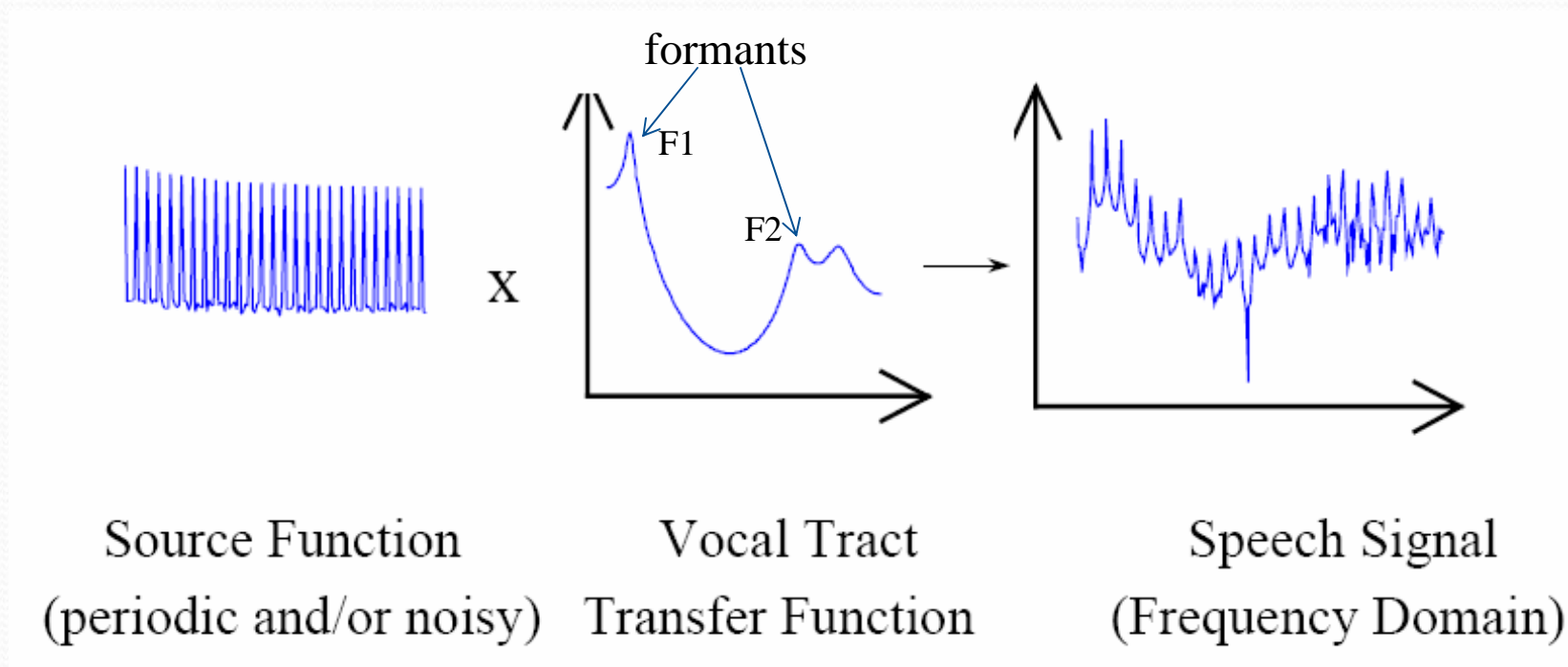
- The Task: to **accurately** convert an acoustic signal into a word sequence, **independent** of speakers and environments
- The Challenges:
  - Phonetic variability: co-articulations, accent, ...
  - Speaker variability: speaking style, vocal tract length, speaking rate, emotion, ...
  - Environment variability: noise, microphone, telephone, ...
- The Applications: dialogue systems, voice search, speech translation, tutoring, ...

# Introduction: ASR system

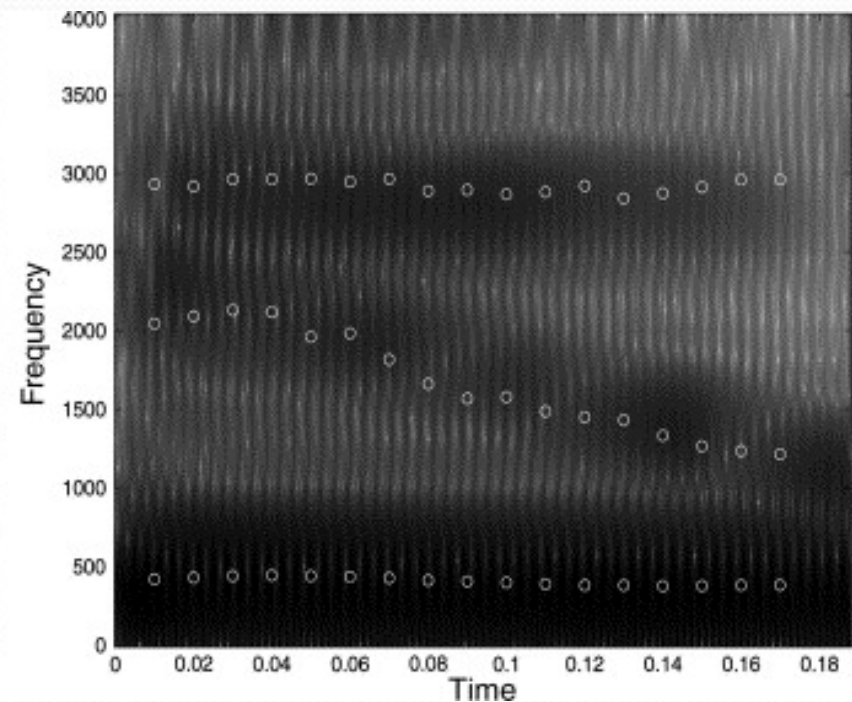
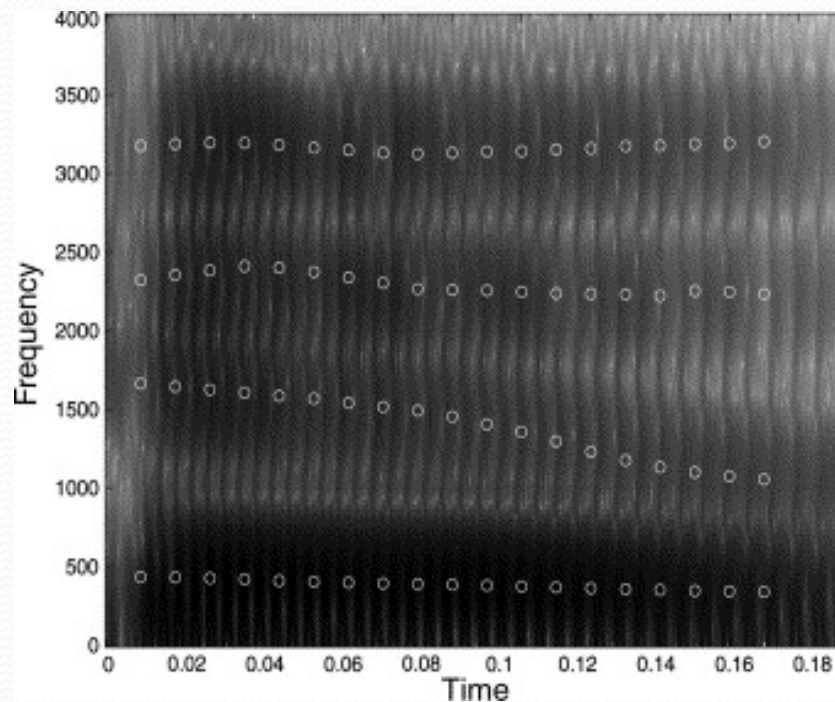


# Introduction: Speech Production

- LTI model: a source signal goes through a vocal tract, producing a speech signal



# Introduction: Spectrogram comparison

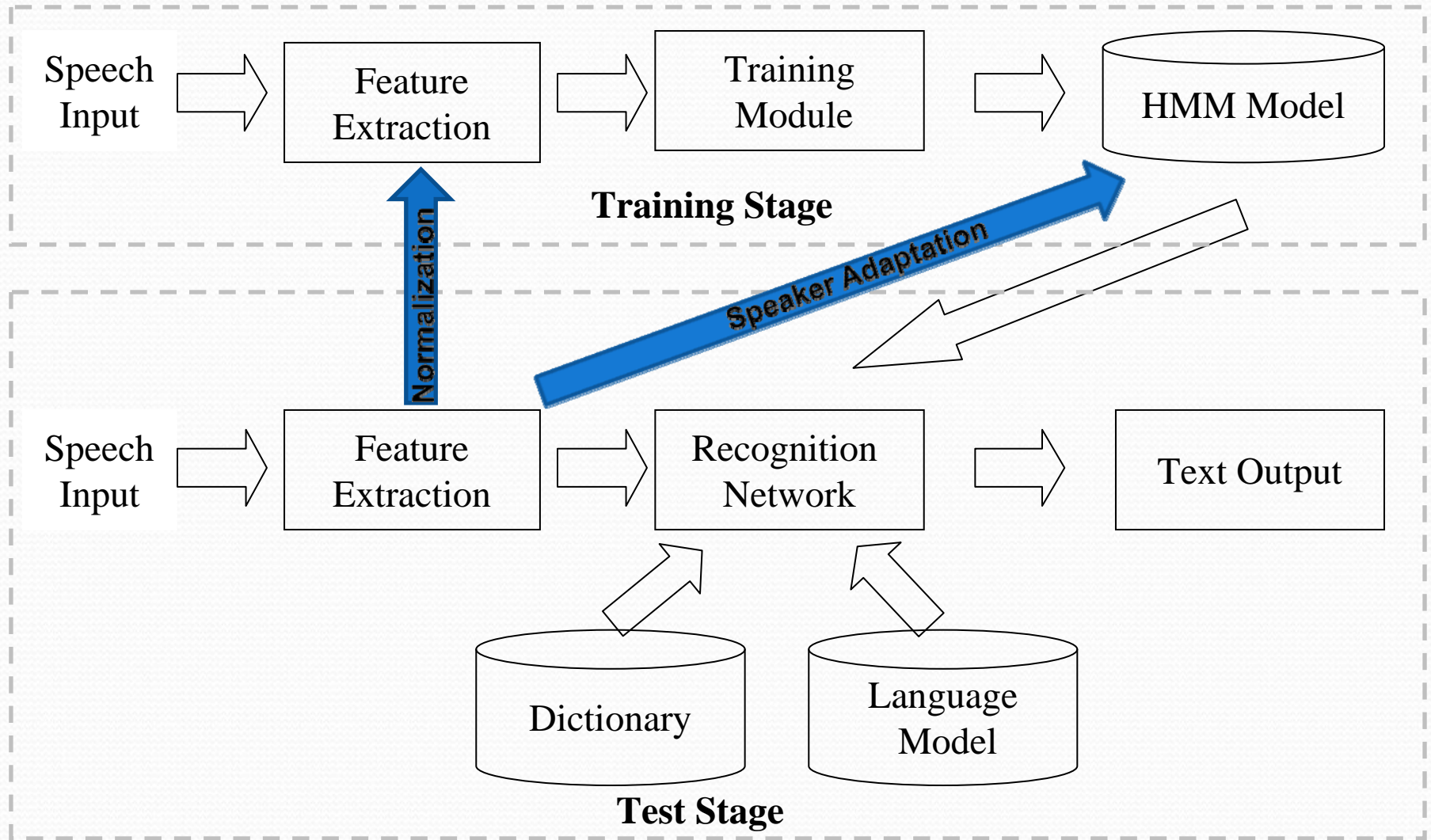


Spectrograms for the sound /uw/ in digit “two” from an adult male speaker (left) and a child speaker (right)

# Introduction: Speaker Normalization

- Cross-speaker robustness is a major issue for ASR
- Inter-speaker variability causes ASR's performance to vary significantly from speaker to speaker
- To reduce inter-speaker variability:
  - Speaker adaptation: tune acoustic models to a specific test speaker (e.g., MLLR, MAP)
  - Speaker normalization: transform test features to match training data through frequency warping(e.g. VTLN)

# Introduction: ASR system

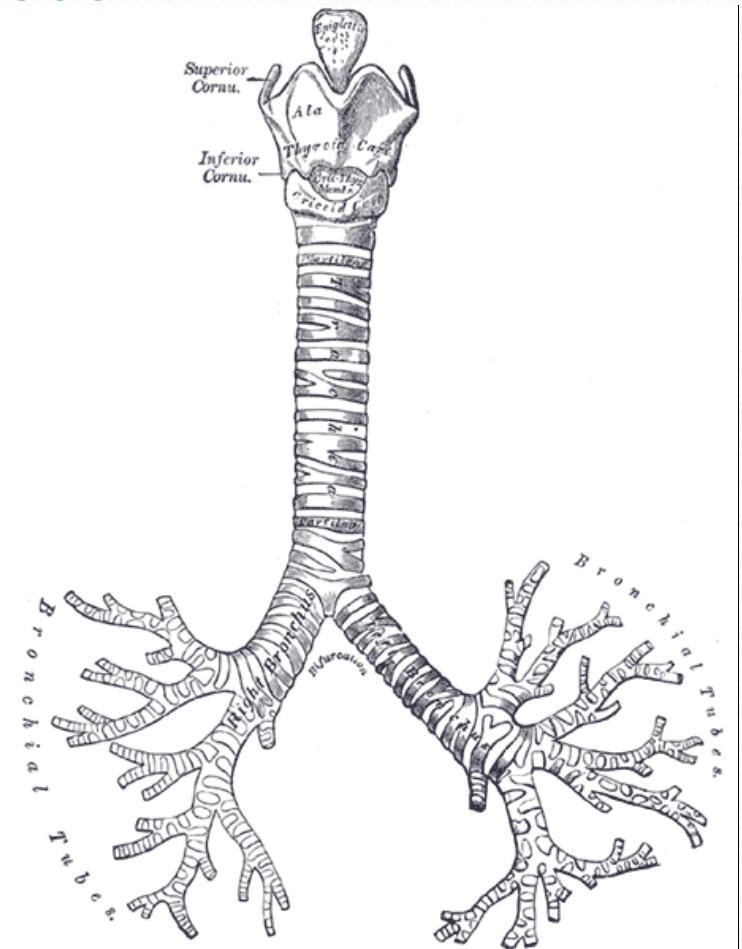


# Limited Data Challenge

- Typical speaker adaptation and normalization methods require enough data to be effective, because they apply statistical parameter estimation
- To lessen the dependency on the amount of data:
  - Apply knowledge-based information (e.g., formants normalization, **subglottal resonance normalization**)

# The subglottal system

- The acoustic system below the glottis consists of the trachea, bronchial, and lungs.
- Similar to the vocal tract, the subglottal system is characterized by poles and zeros. The poles are referred to as subglottal resonances (Sg).

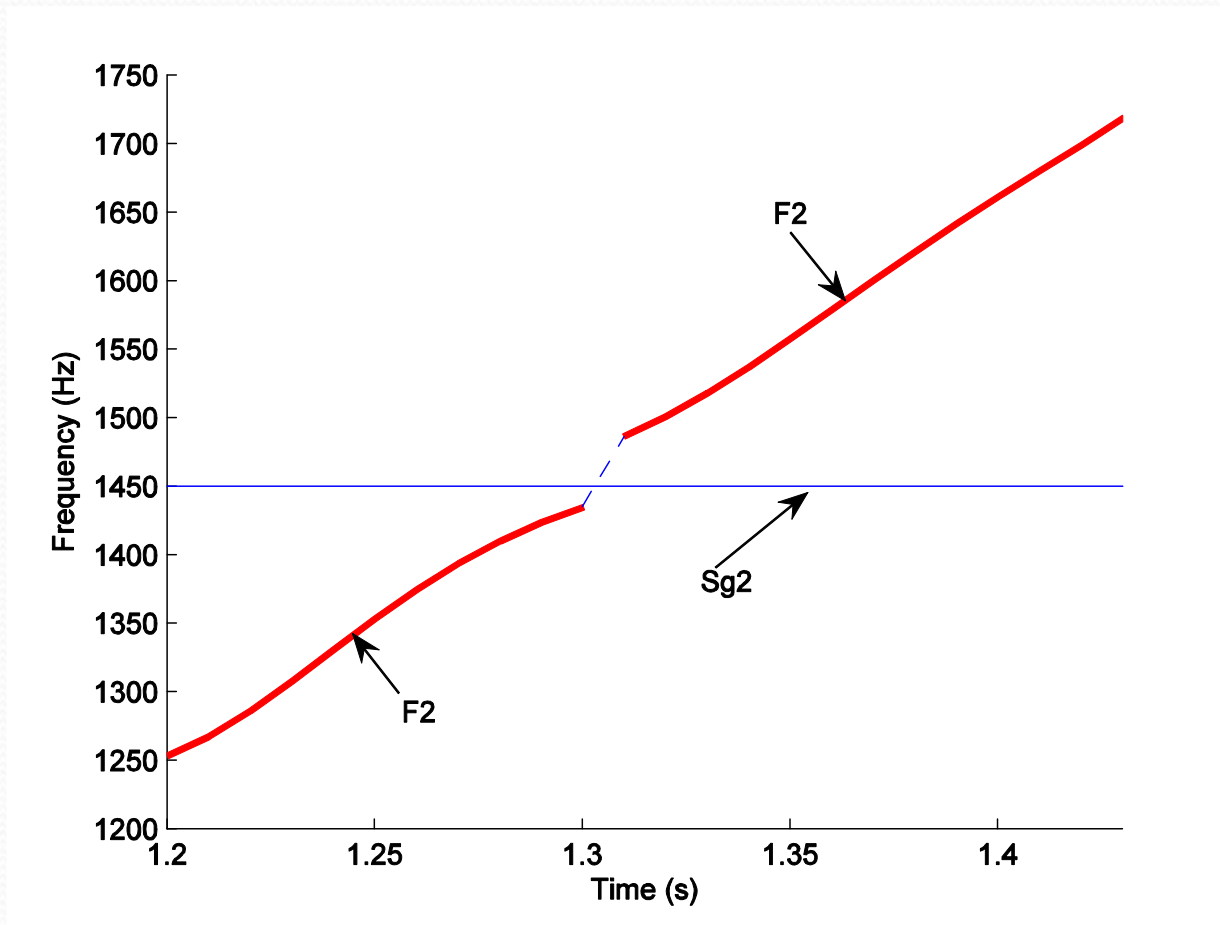


From Gray (1918)

# Coupling of subglottal system

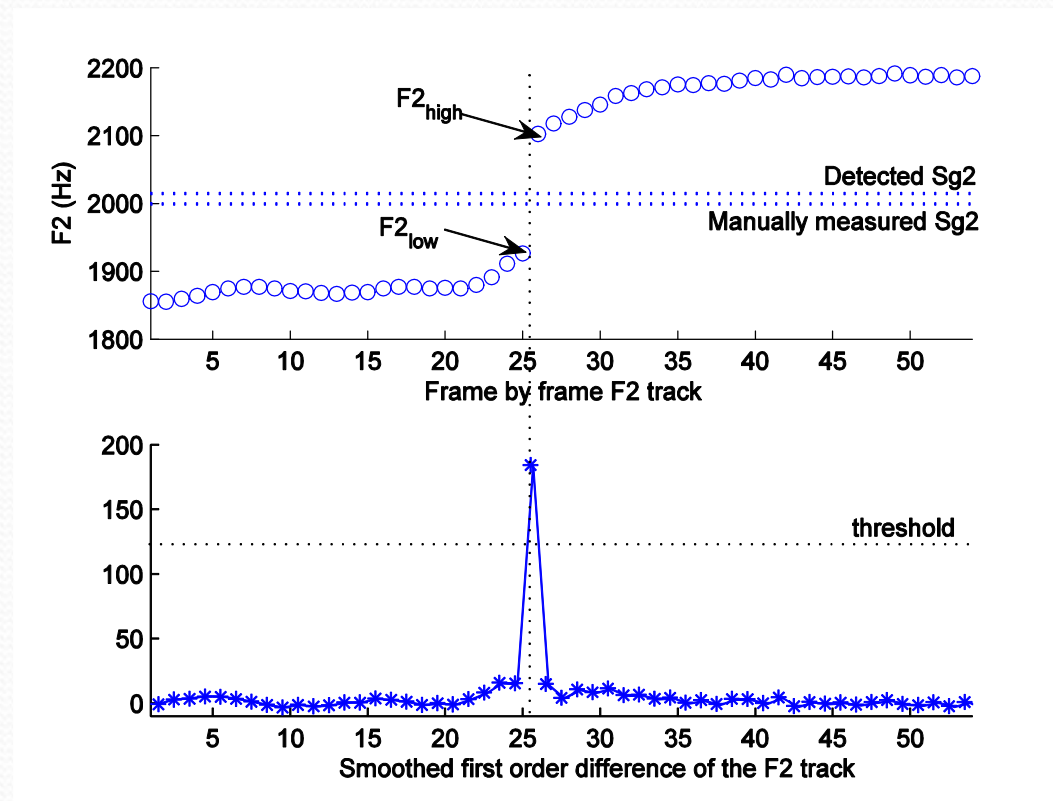
- Introduces pole-zero pairs in the vocal tract transfer function
- Causes the formants to be discontinuous in frequency

# Illustration of F2 track discontinuity



The solid line corresponds to the most prominent spectral peak, which has a jump in frequency when F2 is crossing Sg2.

# Automatic detection of Sg2



# Subglottal resonances

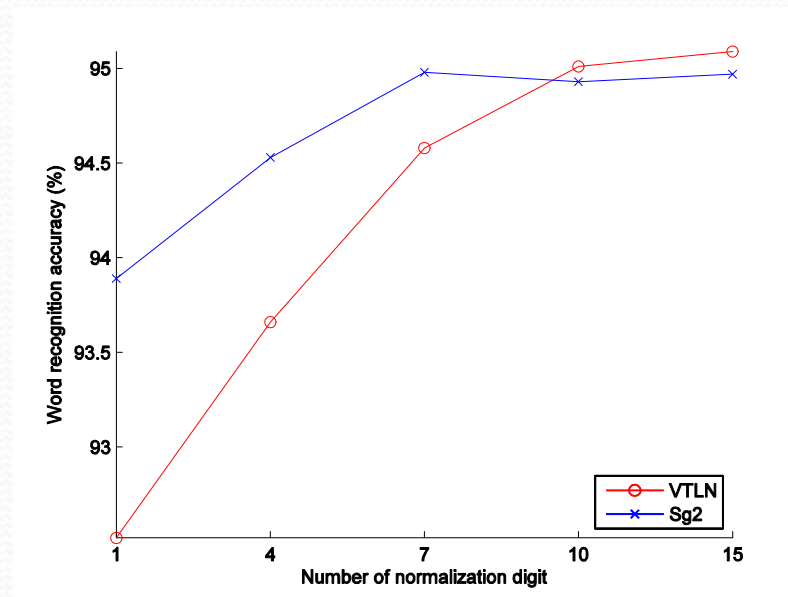
- Theoretically, subglottal resonances remain constant for a given speaker, independent of speech sounds and spoken languages

# Implications of Sg2 invariability

- Sg2 is content-independent → Speaker normalization using Sg2 will be robust for various amounts of adaptation data → Suitable for limited data normalization
- Sg2 is language-independent → Cross-language adaptation based on Sg2 is possible → Useful in ASR applications for second-language learning

# Experimental results: on TIDIGITS

- Acoustic models trained on adults and tested on children's speech



Word recognition accuracy (%) with normalization data varying from 1 to 15 digits

# Experimental results:cross-language

- On Tball database, children's speech
- Acoustic models trained and tested on English

Method	English adaptation	Spanish adaptation data
VTLN	86.85	82.35
Sg2	86.59	<b>85.97</b>

Word recognition accuracy (%) with three normalization words

# Conclusions

- Sg2 normalization performs better than or comparable to VTLN, especially for limited data
- Sg2 normalization produces more robust results than VTLN when performing cross-language adaptation.

# References

- [1] M.J.F. Gales, "Maximum likelihood linear transformations for HMM based speech recognition," *Computer Speech and Language*, vol. 12(2), pp. 75-98, 1998.
- [2] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171-185, 1995.
- [3] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. Speech Audio Processing*, vol. 6(1), pp. 49-60, 1998.
- [4] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. ICASSP*, pp. 346-349, 1996.
- [5] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," in *Proc. ICASSP*, pp. 1039-1042, 1997.
- [6] X. Chi and M. Sonderegger, "Subglottal coupling and its influence on vowel formants," *JASA*, 122(3):1735-1745, 2007.
- [7] S. M. Lulich, "Subglottal resonances and distinctive features," *J. Phon.*, to appear.
- [8] S. Wang, A. Alwan and S. M. Lulich, "Speaker normalization based on subglottal resonances," in *Proc. ICASSP*, pp. 4277-4280, 2008.
- [9] S. Wang, S.M. Lulich, and A. Alwan, "A reliable technique for detecting the second subglottal resonance and its use in cross-language speaker adaptation," in *Proc. Interspeech*, pp. 1717-1720, 2008.
- [10] A. Kazemzadeh, et al, "TBall data collection: the making of a young children's speech corpus," in *Proc. Eurospeech*, pp. 1581-1584, 2005.



Thank you