

- Please view these slides in screen show mode, as they contain animation.

Empowering Text-to-Speech

Ellen Eide, Andy Aaron, Raimo Bakis, Raul Fernandez, Wael Hamza, Michael Picheny, Zhi Wei Shuang*

IBM TJ Watson Research Center
Yorktown Heights, NY USA

* IBM China Research Lab
Beijing, China

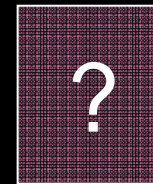
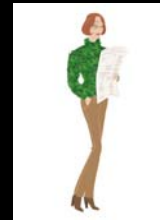


Empower

Basic Text-to-Speech Quality





- First impressions are hard to overcome ...
- Several voices available



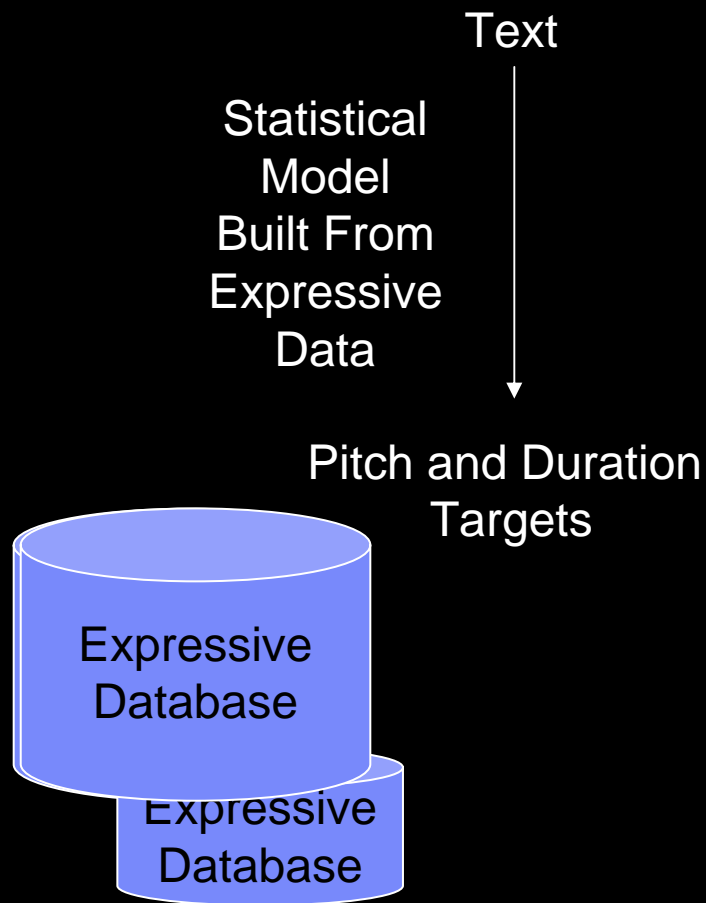
- Quality is text-dependent and speaker-dependent
- Live demo
www.research.ibm.com/tts

Limitations of Basic System

- Produces newsy, slightly perky speech in the speaker's voice 
- Inappropriate for very good or very bad news 
- Solution: expressive TTS
- Cannot satisfy customer's desired voice characteristics
- Solution: voice conversion









Approaches to Generating Expressive Speech

Database Approach



- Collect data to create a search set of data of expressive styles
- Synthesize speech from the expressive database
- Augment neutral segments with expressive ones; reward expressive segments
- Used this approach for good news, bad news, questions, emphasis

Expressive Examples...

	Baseline	Expressive
Good news		
Bad news		
Asking a question		
Showing emphasis		

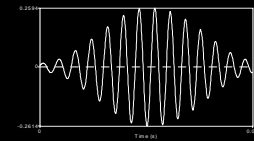
Voice Conversion – Overview of Approach

- Multidimensional problem encompassing pitch, vocal tract resonances, accent, prosody, speaking rate, vocabulary, ...
- Requires the ability to adjust the effect of the pitch and vocal tract (formants) independently.
 - Simple time scaling is insufficient; pitch and formants are uniformly scaled
- Use a decomposition based on sinusoidal analysis to separate pitch and spectral envelope
- Works best when source and target are similar
 - Suggests having a set of source voices which cover the acoustic space ...
 - chose closest to target as a starting point

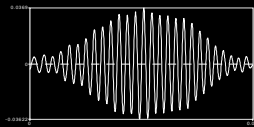
Harmonic Sinusoidal Model of Voiced Speech

$$x(t) = \sum_k A_k \cos(k\omega_0 t + \varphi_k)$$

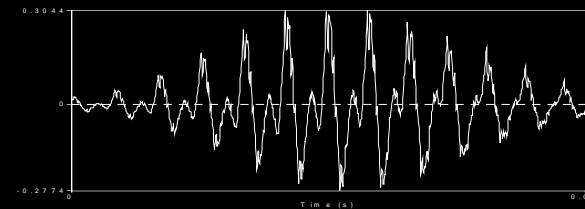
$$A_1 \cos(\omega_0 t + \varphi_1)$$



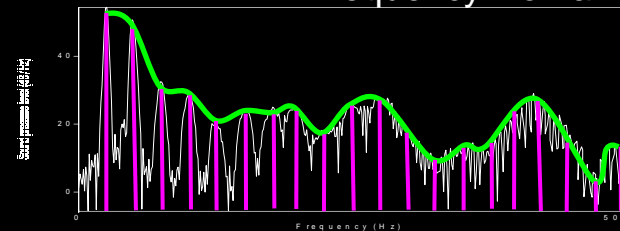
$$A_2 \cos(2\omega_0 t + \varphi_2)$$



Time Domain

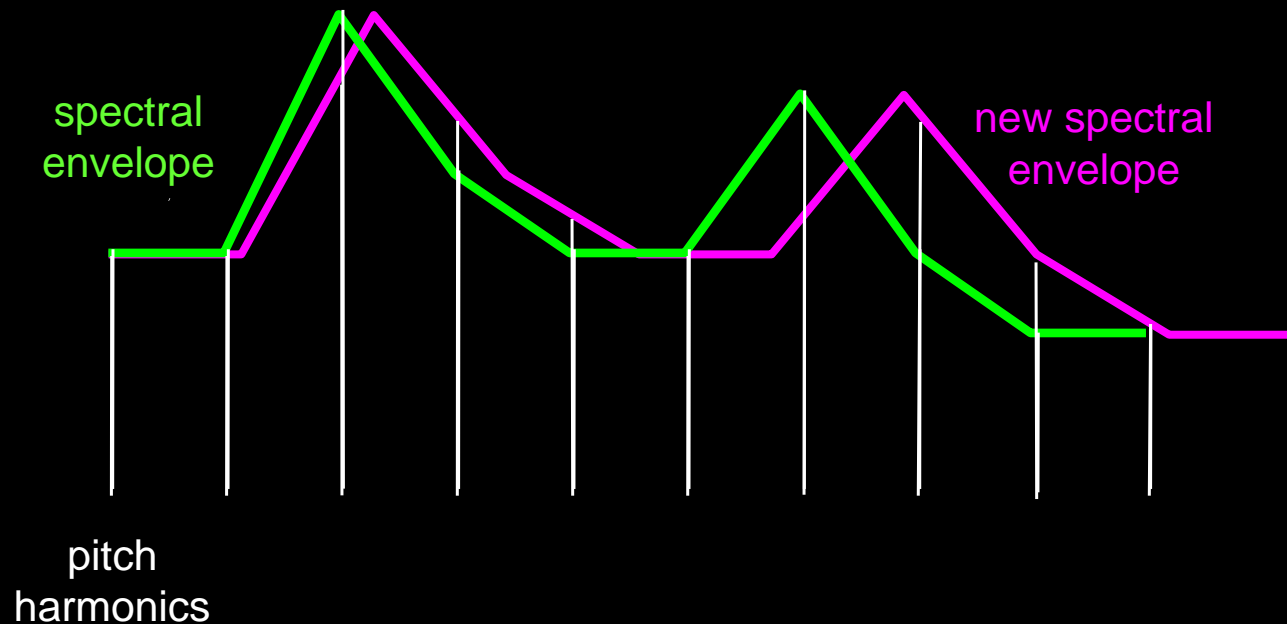


Frequency Domain



Voice Conversion From Sinusoidal Representation

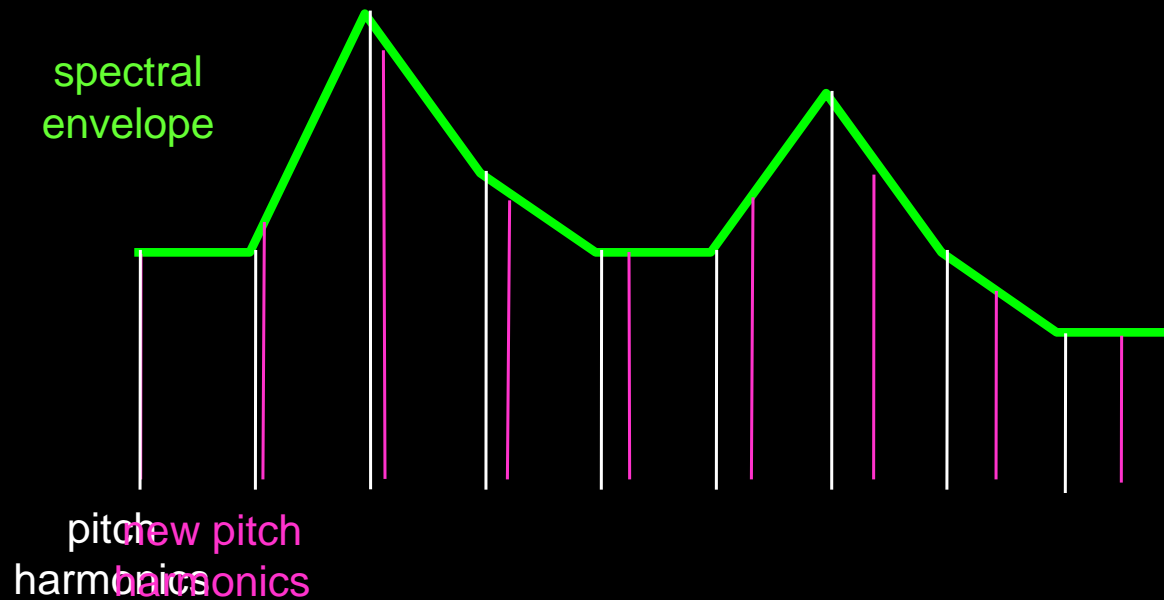
- Once we have the set of amplitudes and phases, we can estimate the spectral envelope, e.g. using linear interpolation between harmonics



- We can then adjust the envelope keeping the original pitch

Voice Conversion From Sinusoidal Representation

- We could also re-sample the original envelope at a new pitch



- Sum new pitch harmonics for reconstruction

$$x_{conv}(t) = \sum_k \tilde{A}_k \cos(k\tilde{\omega}_0 t + \tilde{\varphi}_k)$$

Voice Conversion – Example

Original Speaker



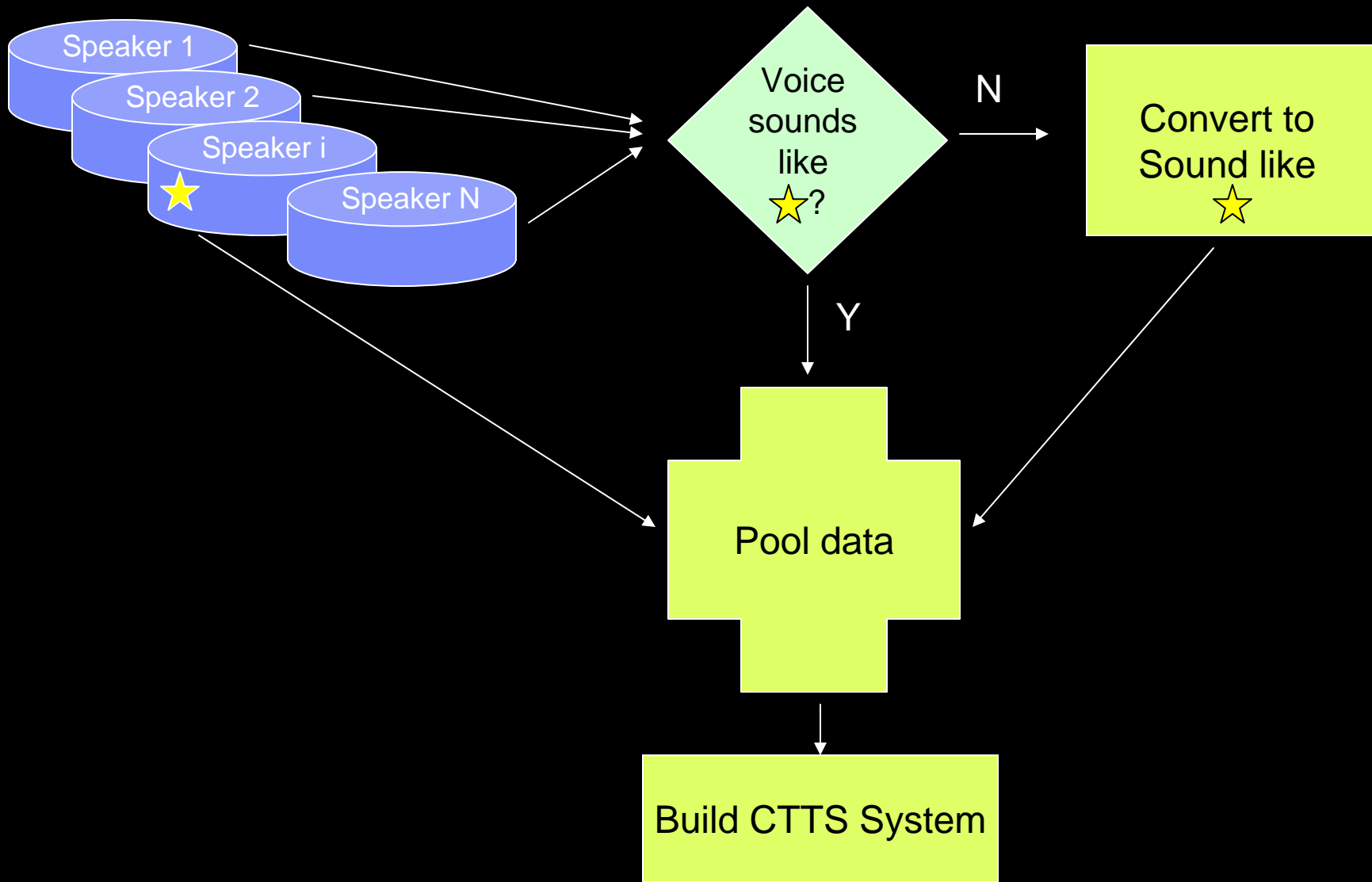
Target



Converted

TTS Customization

- Augment a corpora of customer prompts with off-the-shelf voice data
- Off-the-shelf data can be optionally morphed to better match customer voice
- Use the expressive framework approach for synthesis, treating speakers as styles
- Tune cost matrix to ensure that output speech sounds like target speaker



TTS Customization -- Examples

- Target (40 minutes of speech)
- Auxiliary speaker: off-the-shelf voice (several hours of speech, unmorphed)
- Combined (roughly 50% target, 50% auxiliary)
- Combined (roughly 75% target, 25% auxiliary)



Summary

- Baseline TTS can achieve high quality speech, but has limitations:

One-size-fits-all expression

Fixed voice

- Added expressiveness by recording (small) expressive databases
- Added voice conversion making use of a sinusoidal model representation
- Expressive framework + voice conversion can be used for augmenting a custom dataset

Empowering Your customers and employees with speech technologies

SpeechTEK | 2006
The Voice Solutions Showcase

Empower

Thank You