

Intra-lingual and Cross-lingual Prosody Modelling

PhD Thesis Defense

Gopala Krishna Anumanchipalli

Thesis Committee

Alan W Black, LTI (Chair)

Luís C. Oliveira, IST (Chair)

Justine Cassell, HCII

Mário Figueiredo, IST

Bhiksha Raj, LTI

Isabel Trancoso, IST

Paul Taylor, Google

Speech Translation: PT-STAR*

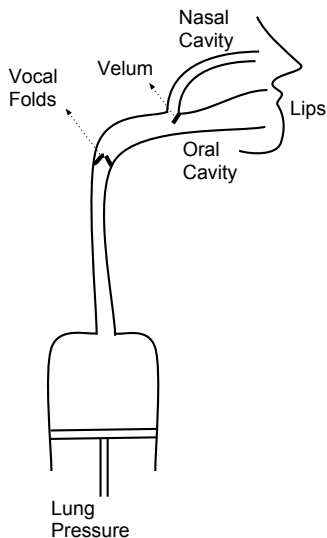
- Aim: Automatic Speech Translation for English \leftrightarrow Portuguese
 - Speech Recognition \rightarrow Language Translation \rightarrow Speech Synthesis
- “Complete” translation of speech input in source language
 - Speaker Identity
 - Sentence Translation
 - Speaker Intent
- This work : Text-to-Speech Synthesis

*Funded by the Fundação para a Ciência e a Tecnologia (FCT), Portugal

Text-to-Speech Synthesis

- Aim: Make computers synthesize speech output from text input
- Desirables
 - Intelligibility
 - Naturalness
 - Flexibility
 - Similarity to a target speaker
 - Robustness

Units of Human Speech



Vowels: uw iy aa ay oy ow

Consonants: k g ch jh t ...

Phonemes:

Eg.: /m ae s ah ch y uw s eh t s/

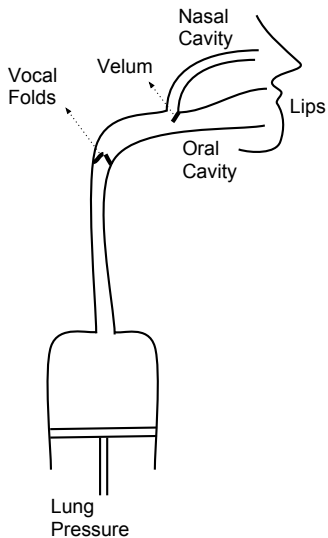
Syllables:

Eg.: /Ma/ /ssa/ /chu/ /setts/

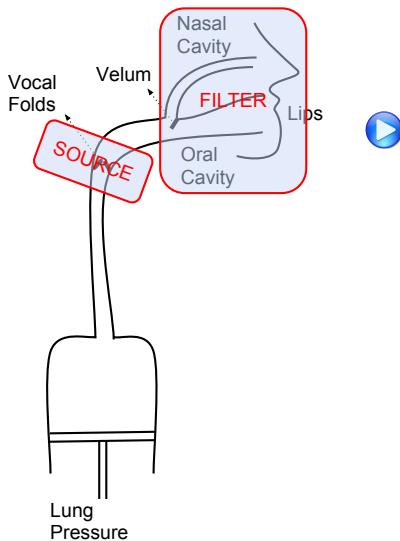
Words:

Eg.: Massachusetts

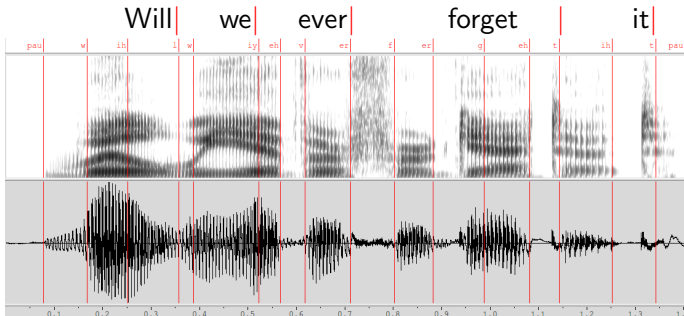
Human Speech Production



Source-Filter Model of Speech

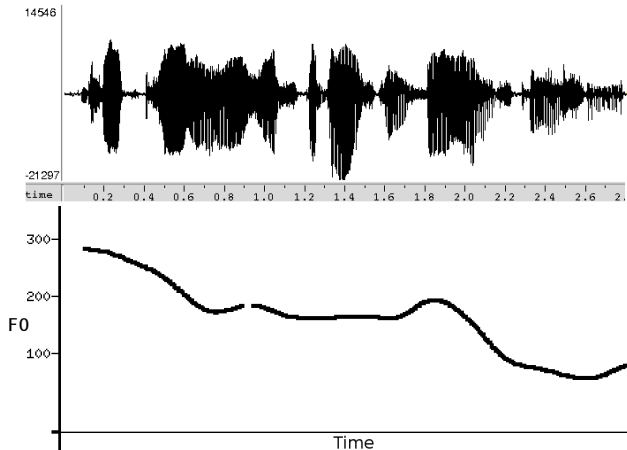


Filter: Spectrum

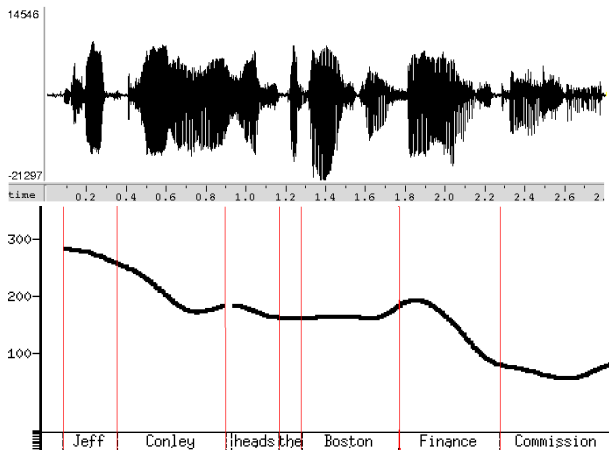


- High density of information (5-10 millisecond *frames* of speech)
- Contributes to Phonetic discriminability between phonemes

Source: Fundamental Frequency



Source: Intonation



Intonation

- The fundamental frequency of vibration of vocal folds
- Systematically conveys the underlying linguistic information
 - Adds Expression, Emotional state, Attitude, Style ...
- Aspects of intonation —
 - Average pitch
 - Pitch range
 - Pitch accents
 - Phrase boundaries

Statistical Parametric Speech Synthesis (SPSS)

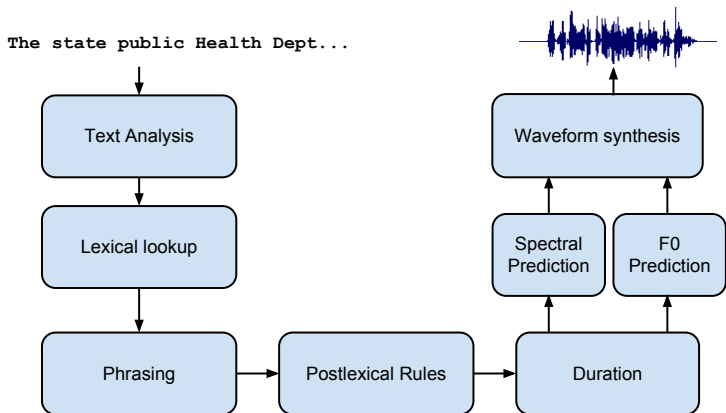


Figure: Runtime architecture: Statistical Parametric Speech Synthesis (SPSS)

CART models in SPSS

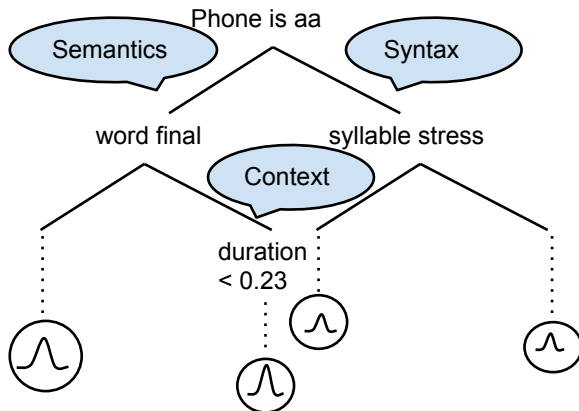


Figure: Decision trees for frame-level statistics of F_0 , Spectra, Duration

Evaluating Synthetic F_0

- Average Pitch (μ)

$$F_0^\mu = \frac{1}{n} \sum_{i=1}^n F_0(i)$$

- Pitch range (σ)

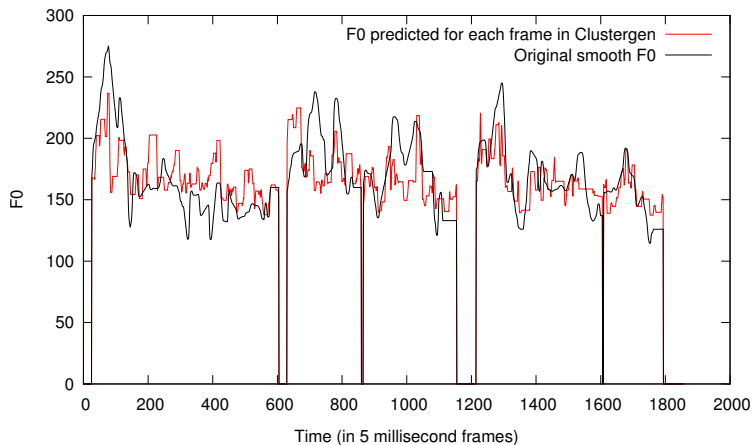
$$F_0^\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (F_0(i) - F_0^\mu)^2}$$

- Correlation against reference

$$r(F_{0X}, F_{0Y}) = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{F_{0X}(i) - F_{0X}^\mu}{F_{0X}^\sigma} \right) \left(\frac{F_{0Y}(i) - F_{0Y}^\mu}{F_{0Y}^\sigma} \right)$$

- Perceptual evaluation by humans is the most reliable measure

Synthetic F_0 Vs. Natural F_0



	F_0^μ	F_0^σ	
Natural F_0	167.852	30.276	$r=0.49$
Synthetic F_0	168.673	18.549	

Issues with synthetic F0 in SPSS

- Model not capturing the intonation phenomenon
- Features not discriminative enough to explain F_0 variance
- Text \rightarrow F0 relationship ill-modelled at the frame level
 - Loss of Naturalness
 - Loss of Expression
 - Loss of Variance

Thesis Statement

It is possible to computationally model intonation, through —

Thesis Statement

It is possible to computationally model intonation, through —

- *A statistical framework for expressive **modelling***

Thesis Statement

It is possible to computationally model intonation, through —

- *A statistical framework for expressive **modelling***
- ***Conversion** between speakers within a language*

Thesis Statement

It is possible to computationally model intonation, through —

- *A statistical framework for expressive **modelling***
- ***Conversion** between speakers within a language*
- *Cross-lingual **transfer** for improving speech translation*

Thesis Statement

It is possible to computationally model intonation, through —

- *A statistical framework for expressive **modelling***
- ***Conversion** between speakers within a language*
- *Cross-lingual **transfer** for improving speech translation*
- ***Conformity** with existing theoretical frameworks of intonation*

Contributions of this Thesis

- A Phonologically sound modelling unit for F_0 in SPSS (Anumanchipalli '13a, Sitaram '13)
- A Multi-tier architecture for F_0 synthesis (Anumanchipalli '11)
- Style-aware F_0 transformation in voice conversion (Anumanchipalli '13b)
- Intent transfer in speech translation (Anumanchipalli '12b)

Intonation Modelling

Paradigms in Intonation

Several theories are proposed previously to explain Intonation

- Physiological (Fujisaki '83, van Santen '05, Bailly and Holm '05 ...)
- Phonological (Lieberman '77, Pierrehumbert '80, Silverman '92 ...)
- Stylization ('t Hart and Collier '73, Hirst '93, Taylor '00 ...)

1. Physiological Fujisaki Model

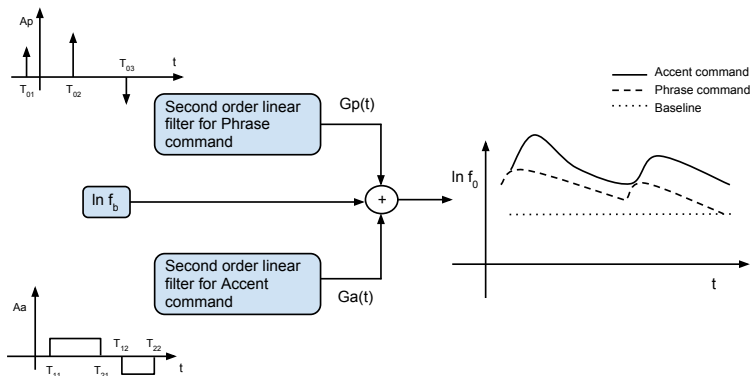


Figure: Fujisaki model — $\ln (F_0)$ as a superposition of the baseline, phrase and accents.

Fujisaki Model Summary

- Physiologically motivated
- Strict assumptions on *phrase* and *accent* shapes
- Postulates additive *Global* and *Local* components
 - Potential to better explain variance ✓

2. Phonological Tone Sequence -ToBI

F_0 as a finite number of **Tones** and **Break Indices** (Silverman '92)

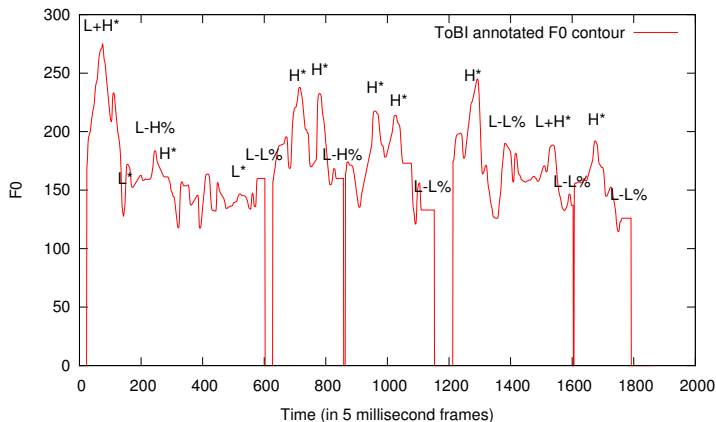


Figure: A natural F_0 contour annotated with ToBI labels

Tone Sequence Model Summary

- Sequential tonal structure
- Distinction between Pitch Accent and stress
- Qualitative description
- Postulates **limited** number of descriptive shapes
 - Suitable to cluster in SPSS ✓

3. Stylization: TILT model of Intonation

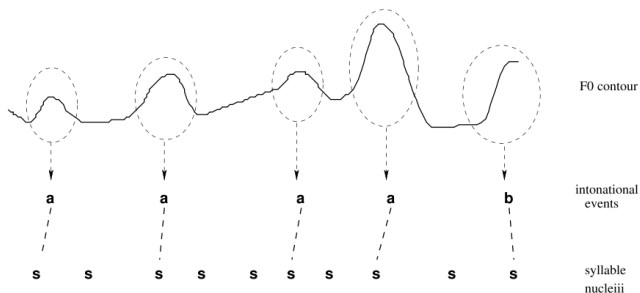


Figure: F_0 contour as a sequence of Intonational Events (Taylor '00)

Tilt Analysis of F_0

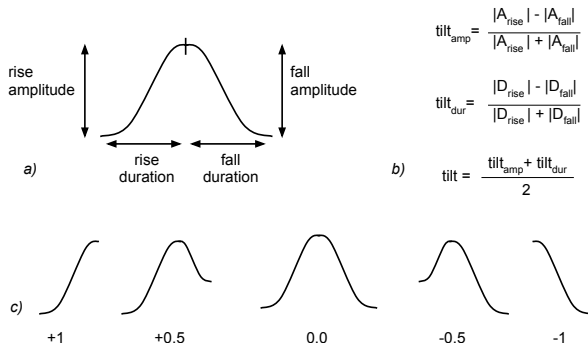


Figure: a) Analyzing each intonational event using its rise/fall values. b) three parameters to code any arbitrary rise/fall event c) Examples of 5 pitch accents with the continuous *tilt* value ranging from -1 to +1

TILT Model Summary

- Realize the F_0 curve as a series of **syllable-level** shapes
- Associate the event with its syllable
- Quantify each Pitch Accent in terms of Tilt
- A structured representation, that can be learned from data ✓

Where is the Pitch Accent ?

Which phonological level to anchor accent shapes to ?

- Phrase
- Word
- Syllable
- Phoneme

Distribution of Phonological units

Table: Distribution of phonological units in 1 hour of Radio news speech

Unit	Number of instances
Sentence	464
Phrase	1052
Word	9214
Syllable	14717
Phoneme	38523
Phoneme state	115569
Frame	592830

- Heavily skewed towards shorter units
- Higher questions are overwhelmed by lower positional features

Comparing F_0 Models at different levels of phonology

- Models built with comparable, appropriate question sets
- All other parts of the TTS remain the same

Table: Objective comparison of original and synthesized for F_0

Modeling unit	F_0	
	Mean	Std/dev
Original	167.85	30.28
Frame Predicted	168.67	18.55
Syllable Predicted	175.25	16.48
Word Predicted	177.00	18.95

Objective Comparison: RMSE & Correlation

Table: Objective comparison of predicted F_0 contours against references

Modeling unit	Predicted F_0	
	RMSE	CORR
Frame	28.02	0.49
Syllable	30.33	0.40
Word	30.34	0.44

Qualitative Analysis: Frame Predicted F_0 contour

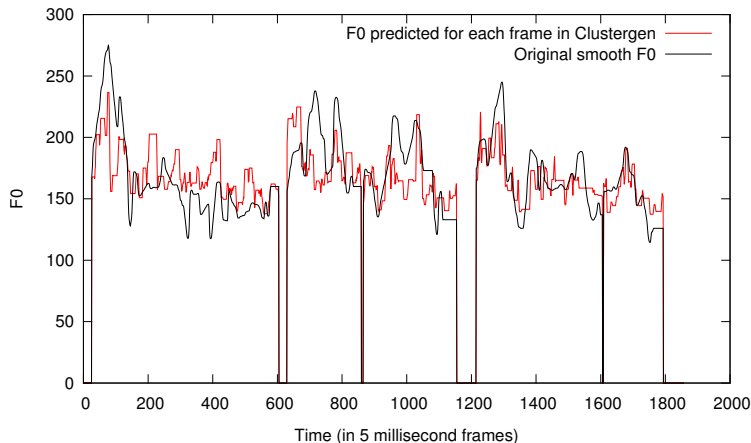


Figure: Prediction of F_0 at the frame level. **Naturalness lost.**

Qualitative Analysis: Syllable Predicted F_0 contour

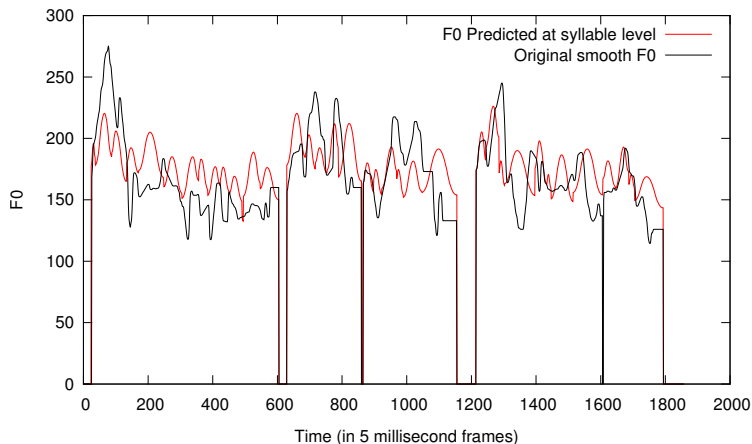


Figure: Prediction of F_0 at the Syllable level. Too many peaks.

Qualitative Analysis: Word Predicted F_0 contour

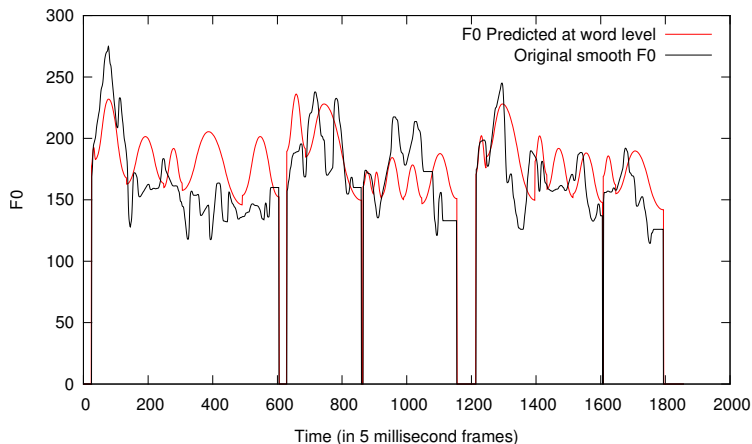
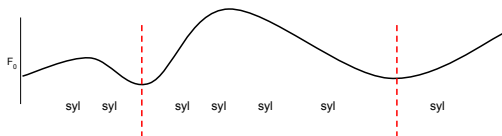


Figure: Prediction of F_0 at the Word level. **Too few peaks.**

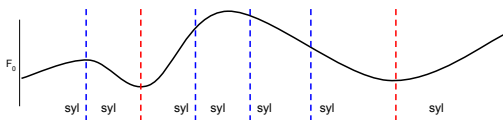
Proposal : Accent Groups (Anumanchipalli et al., '13a)



Groups of syllables which have only one accent on them

- May be spread across words
- May consist of only one syllable
- Bounded by intermediate phrase boundaries
- End in prosodic phrase boundaries
- Task/ speaker/ language/ dialect dependent

Accent Group Discovery from Speech



- A one-pass reconstruction strategy linear in $\#$ Syllables
- Parametrize each syllable under a Tilt Parametrization
- Group syllables together reconstruction error gain
- Parametrize each such Accent Group and Reconstruct F_0

Accent Group Discovery from Speech

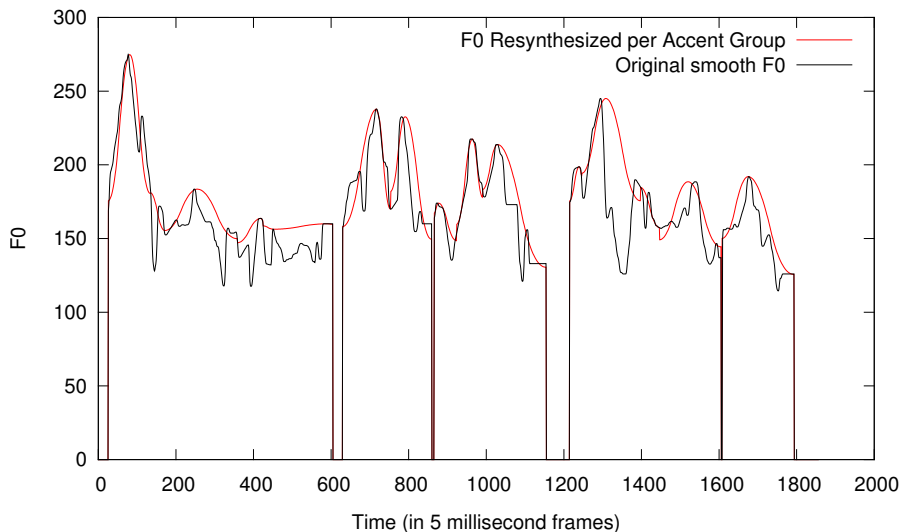


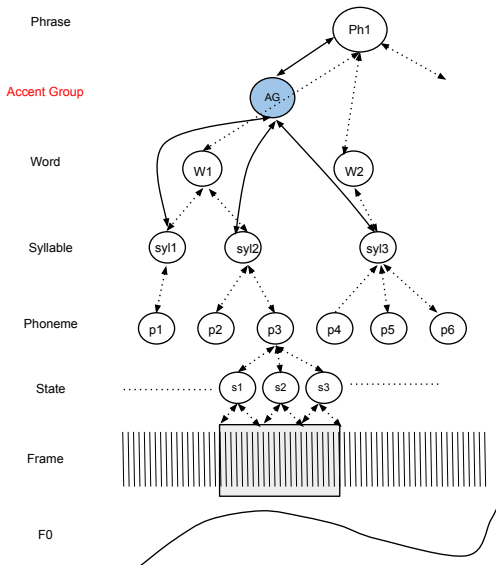
Figure: Resynthesized F_0 for automatically detected Accent Groups

Accent Groups against other phonological units

Table: Accent Groups against other phonological units

Unit	Number of instances
Sentence	464
Phrase	1052
Accent Group	7751
Word	9214
Syllable	14717
Phoneme	38523

Integration with Festival Prosodic Structure



Prediction of Accent Groups and F_0 from Text

- Train Grammars of Accent Group parses in training data
- Predictive models using grammatical and linguistic context
- Decision at each syllable if an Accent Group boundary follows
- Predict pitch accent shape for each Accent Group
 - Question set of 83 (richer semantic features selected in CART)

F_0 Prediction with Accent Groups

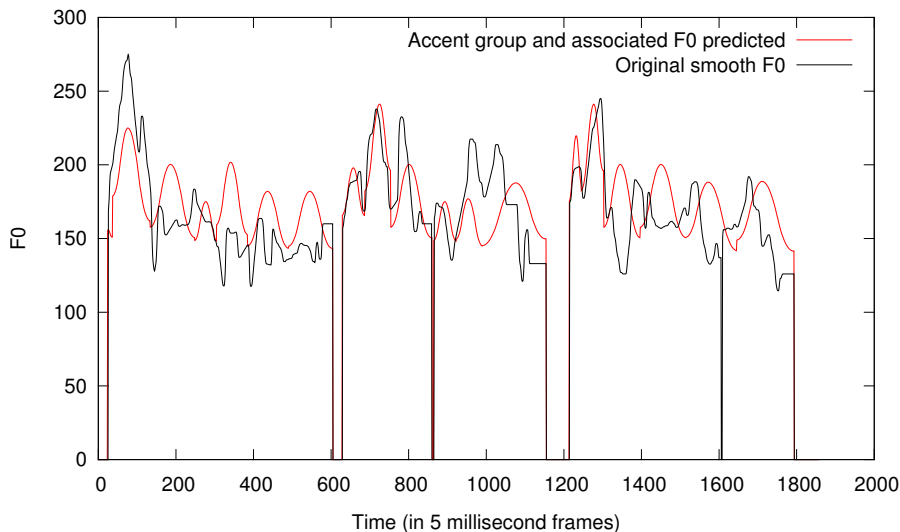
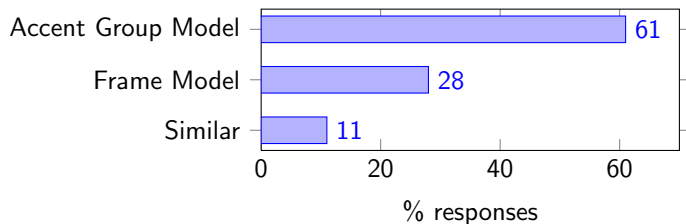


Figure: Predicted F_0 for predicted Accent Groups

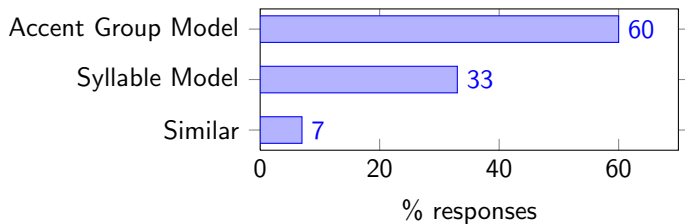
Subjective Judgment[†]: Frame Vs. Accent Group

- Compared against SPSS at Frame level without SPAM



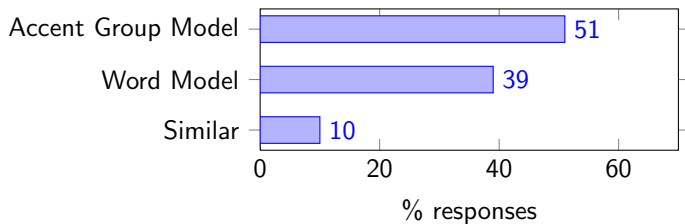
[†]50 native listeners on the Amazon Mechanical Turk

Subjective Judgment[‡]: Syllable Vs. Accent Group



[‡]50 native listeners on the Amazon Mechanical Turk

Subjective Judgment[§]: Word Vs. Accent Group



[§]50 native listeners on the Amazon Mechanical Turk

Improved Variance

Table: Comparing Accent Group against other units for synthetic F_0

Modeling unit	F_0	
	Mean	Standard Deviation
Original	167.85	30.28
Frame Predicted	168.67	18.55
Syllable Predicted	175.25	16.48
Accent Group	173.08	21.24
Word Predicted	177.00	18.95

Objective comparison over tasks and modeling units

Unit	SLT (read)		F2B (news)		TATS (audiobook)	
	err	corr	err	corr	err	corr
Frame	10.97	0.62	37.22	0.38	29.95	0.08
Syllable	12.15	0.47	37.05	0.23	25.28	0.07
Word	12.65	0.46	36.30	0.33	25.80	0.08
Accent Group	13.13	0.43	35.79	0.33	25.96	0.06
Accent Group Oracle	11.49	0.51	35.50	0.34	24.91	0.09

- Read isolated sentences are more predictable than Multi-paragraphs
- Higher phonological units better suited to model expressive F_0
- Given the ideal grouping, Accent groups are the **optimal** unit

Can we do better ?

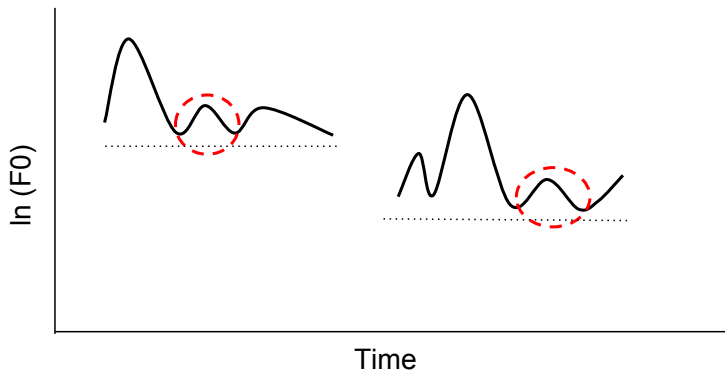


Figure: Equivalent pitch accents from two different phrases. Wrong to Average!

A Multi-tier Additive Model

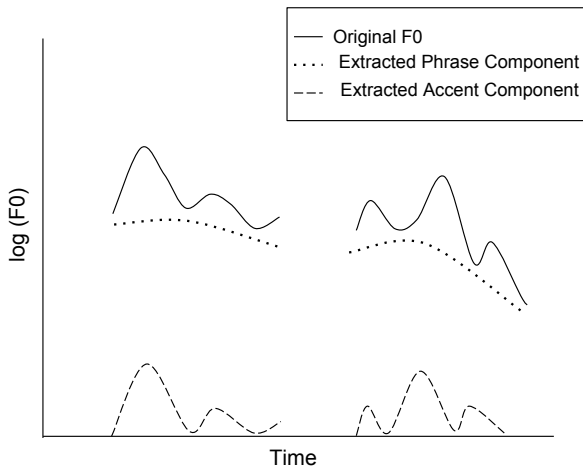


Figure: The notion of a 2-tier architecture for F_0 production.

SPAM: Statistical Phrase/Accent Model (Anumanchipalli et al., '11)

Iterative decomposition and model training

- Initialize Phrase as minimum over each Accent Group
- Parametrize residual as Tilt Accents
- Apply constraints on models for prediction from long/short features
- Reestimate contour and adjust phrase with resynthesis error
- Iterative improvement of Phrase/Accent component Models

RMSE across Training iterations

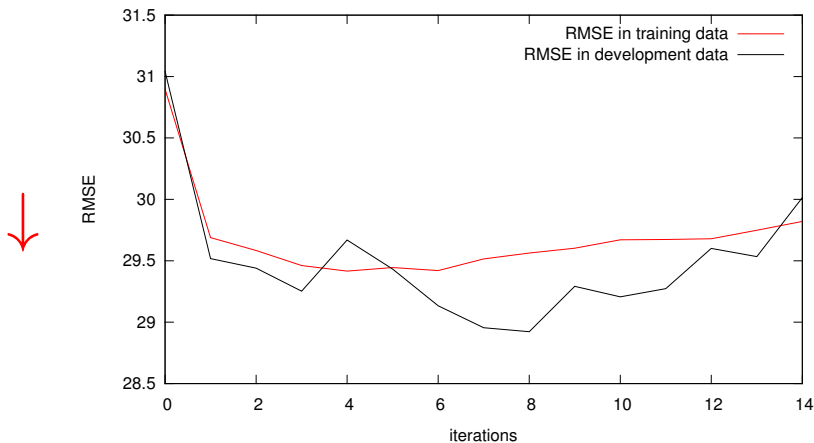


Figure: RMSE on Training and Development sets over iterations

RMSE across Training iterations

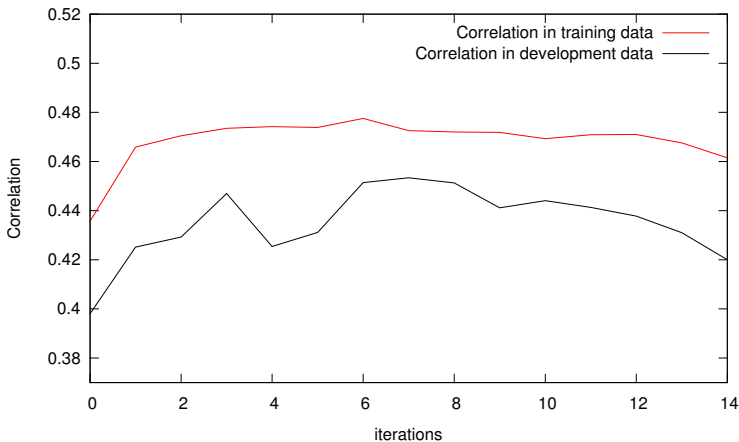


Figure: Correlation on Training and Development sets over iterations

Does SPAM gain from more data ?

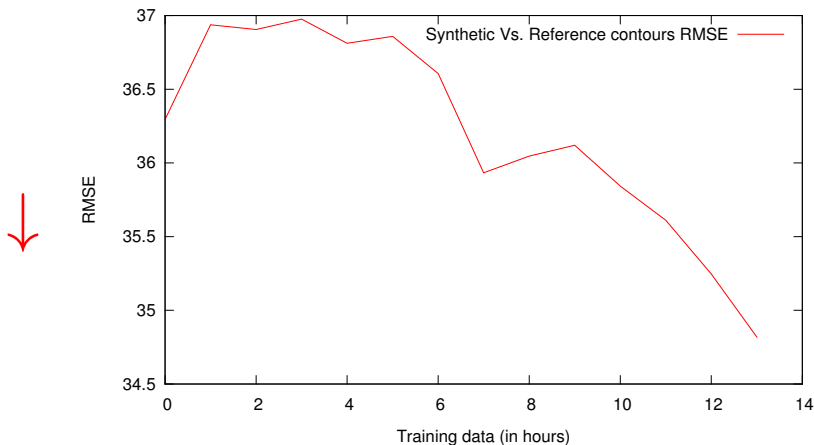


Figure: RMSE of test set with increasing amounts of training data

Does SPAM gain from more data ?

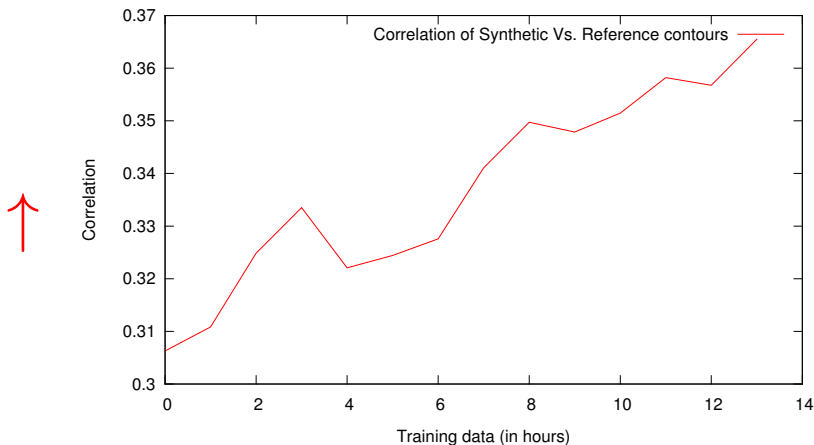


Figure: Correlation on test set using increasing amounts of training data

Models from best iterations

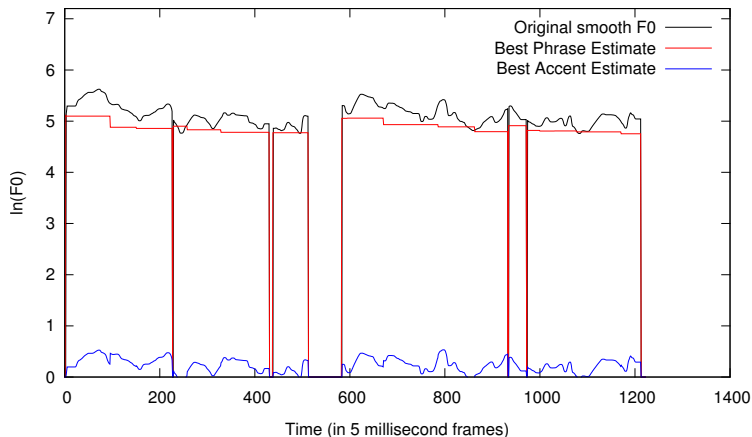


Figure: Best component splits after training

Prediction with SPAM F_0 model

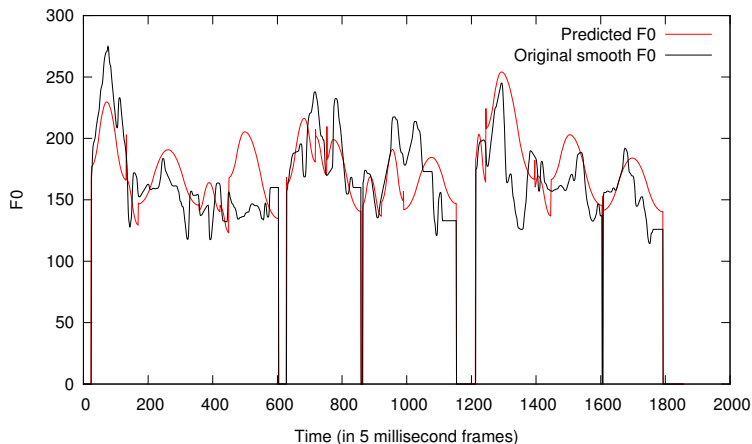


Figure: Prediction of F_0 on an unseen sentence

Objective Comparisons

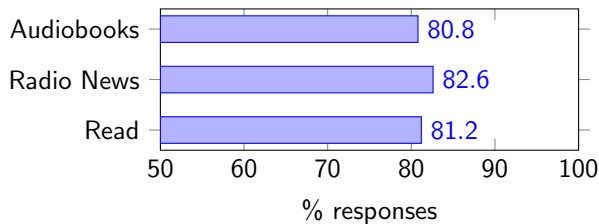
Table: Comparing SPAM against other Modeling units for synthetic F_0

Modeling unit	F_0	
	Mean	Standard Deviation
Original	167.85	30.276
Frame Predicted	168.67	18.55
Syllable Predicted	175.25	16.484
Word Predicted	177.00	18.95
Accent Group	173.08	21.24
Multi-tier Accent Group	168.77	26.41

Table: Objective comparison of synthetic F_0









Model	RMSE	CORR
Multi-tier Accent Group	32.06	0.40
Accent Group Oracle	28.96	0.45

Subjective Preference to SPAM[¶]



[¶]12 Native listeners

Examples Before & After SPAM

Several wealthy and benevolent individuals in the county subscribed largely for the erection of a more convenient building in a better situation.		
Lastly, I saw Mr. Mason was submissive to Mr. Rochester		
O Tágide, perto da Escola de Belas Artes, passo uma vista de olhos pela montra da Livraria Sá da Costa, pouco à frente da secular Bertrand.		
Dirigido a todas as gerações, o Licor Beirão continua presente nas festas portuguesas.		

Summary

- Accent Groups are optimal for modeling Pitch Accents
- A Multi-tier Phonological model for generation of F_0
- A computational framework for training/synthesis using SPAM
- Preserve expression natural variance
- Minimal Task/Language dependence

Future Directions

- Extensions to include microprosody
- Explicitly downgrading certain Accent Groups to connections
- Automatic speaker/dialect characterization
- Other Parametrizations for representing Pitch Accents
- Other Machine Learning models for F_0 modelling

Voice Conversion

Voice Conversion

- Goal: Transform a speaker's speech to sound like a target speaker
- Speaker characteristics
 - Spectrum
 - Speaking Style
 - Voice quality

Speaking style

Professional impersonators capture aspects of speaking style

[Zetterholm, 2006]

- Rhythm
- Intonation
- Stress patterns across words and phrases

The biggest challenge at this stage for voice conversion algorithms is the control (modeling, mapping and modification) of the speaking style of a speaker [Stylianou, 09].

Conventional Intonation Transformation

- Obtain parallel data from the source and target speakers
- Convert source speaker's F0 mean and range to match target

$$F0_{(t)}^{tgt} = \frac{\sigma^{tgt}}{\sigma^{src}} \left(F0_{(t)}^{src} - \mu^{src} \right) + \mu^{tgt}$$

Eqn 1: Z-score transformation approach for F0 conversion

Illustration: American Female \rightarrow American Male

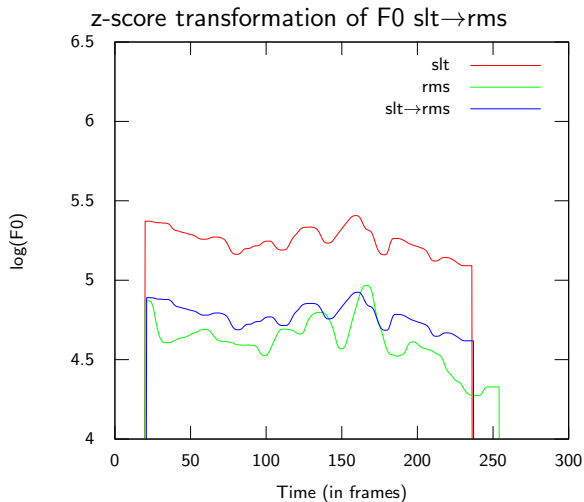
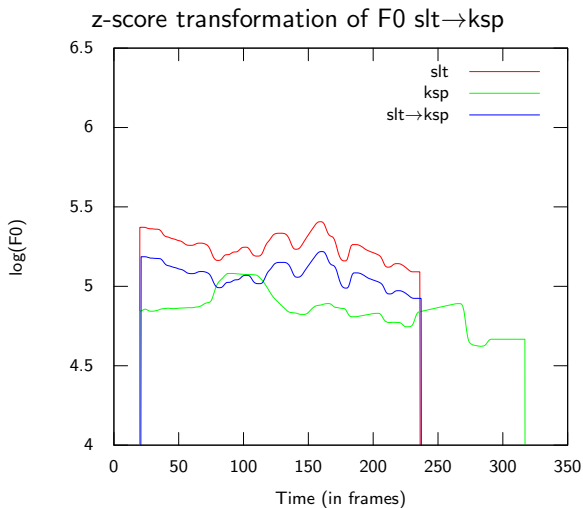


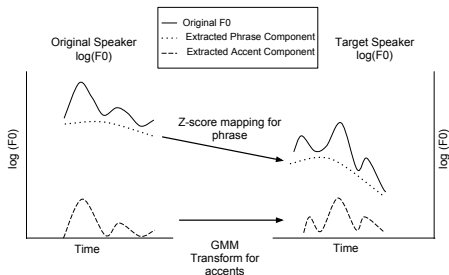
Illustration: American Female \rightarrow Indian Male



Issues with z-score F0 transformation

- Z-score transformation technique converts F0 per each frame (5-10 milliseconds)
- It cannot capture
 - Position of the pitch accents
 - Shape of the pitch accents
 - Utterance specific artefacts
- Doesn't efficiently use the parallel data!

Mapping of Pitch Accents (Anumanchipalli et al., '13b)



- Use SPAM to decompose F_0 contours respective phrase and accent components
- Train Joint density GMMs (Toda '06) of F_0 over Accent Groups
- Compute z-score parameters for transforming phrases
- At test time, apply a z-score transform on phrases and GMM Joint Density Estimation over accent shapes

Style Capturing Intonation Transformation

- Transform phonological regions (Accent Groups) rather than frames

Style Capturing Intonation Transformation

- Transform phonological regions (Accent Groups) rather than frames
- Simulating parallel speech data from the two speakers

Style Capturing Intonation Transformation

- Transform phonological regions (Accent Groups) rather than frames
- Simulating parallel speech data from the two speakers
 - Detect Accent Groups in source speaker's speech

Style Capturing Intonation Transformation

- Transform phonological regions (Accent Groups) rather than frames
- Simulating parallel speech data from the two speakers
 - Detect Accent Groups in source speaker's speech
 - 'Force align' source speaker's Accent Groups on target speaker's F_0

Style Capturing Intonation Transformation

- Transform phonological regions (Accent Groups) rather than frames
- Simulating parallel speech data from the two speakers
 - Detect Accent Groups in source speaker's speech
 - 'Force align' source speaker's Accent Groups on target speaker's F_0
 - Parameterize accent shapes within each Accent Group

Style Capturing Intonation Transformation

- Transform phonological regions (Accent Groups) rather than frames
- Simulating parallel speech data from the two speakers
 - Detect Accent Groups in source speaker's speech
 - 'Force align' source speaker's Accent Groups on target speaker's F_0
 - Parameterize accent shapes within each Accent Group
 - Train source→target pitch accent mapping function

Style Capturing Intonation Transformation

- Transform phonological regions (Accent Groups) rather than frames
- Simulating parallel speech data from the two speakers
 - Detect Accent Groups in source speaker's speech
 - 'Force align' source speaker's Accent Groups on target speaker's F_0
 - Parameterize accent shapes within each Accent Group
 - Train source→target pitch accent mapping function
- Apply mapping function on each pitch accent of the source speaker to predict target speaker's pitch accent

Joint density Modeling of Parallel Pitch Accents

- Joint vectors of source-target Pitch Accents $z_t = \begin{bmatrix} x'_t \\ y'_t \end{bmatrix}$
- Modelled as mixture of M Gaussians

$$P(z_t | \lambda^{(z)}) = \sum_{m=1}^M w_m \mathcal{N}(z_t; \mu_m^{(z)}, \Sigma_m^{(z)})$$

- The covariance matrix $\Sigma_m^{(z)} = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix}$

Estimating the Target Pitch Accent

- Predict most likely y_t , given $\lambda^{(z)}$ and an unseen x_t

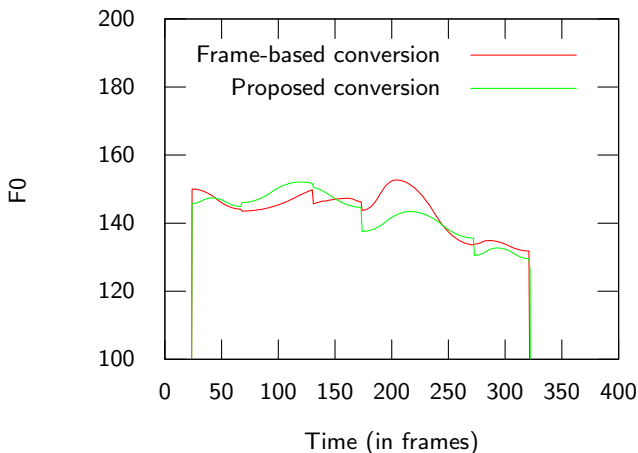
$$\hat{y}_t = \sum_{i=1}^M p(m_i|x(t), \lambda^{(z)}) E(y_t|x_t, m_i, \lambda^{(z)}),$$

$$E(y_t|x_t, m_i, \lambda^{(z)}) = \mu_i^{(y)} + \Sigma_i^{(yx)} \Sigma_i^{(xx)^{-1}} (x_t - \mu_i^{(x)}),$$

$$p(m_i|x(t), \lambda^{(z)}) = \frac{w_i \mathcal{N}(x_t; \mu_i^i, \Sigma_i^{(xx)})}{\sum_{j=1}^M w_j \mathcal{N}(x_t; \mu_j^{(x)}, \Sigma_j^{(xx)})}$$

Movement of the H*

- Tested on several speaker pairs from Arctic databases



An illustration of slt→ksp F0 conversion

Objective Comparisons

Speaker pair	Z-score transform		2-stage conversion	
	RMSE	CORR	RMSE	CORR
bd1-slt	0.49	0.38	0.47	0.52
bd1-ksp	0.26	0.45	0.29	0.53
bd1-awb	0.31	0.53	0.31	0.65
bd1-rms	0.59	0.46	0.42	0.41
ksp-bd1	0.32	0.56	0.31	0.56
ksp-slt	0.47	0.42	0.44	0.51
ksp-rms	0.49	0.34	0.70	0.51
ksp-awb	0.33	0.56	0.30	0.63
rms-bd1	0.22	0.57	0.23	0.59
rms-slt	0.63	0.25	0.44	0.49
slt-bd1	0.64	0.47	0.35	0.50
slt-rms	0.92	0.53	0.48	0.31

Summary

- The approach moves the H^* as appropriate to a target speaker
- Complete conversion needs transforming all aspects of Prosody
- These include - phrasing, duration and Accent Grouping

Speech Translation

Speech-to-Speech Machine Translation

Goal: Convert speech input to speech output in another language

Traditional Serial Approach:

- Recognize source speech (ASR)
- Translate ASR hypothesis to target language (SMT)
- Synthesize translated sentence in target language (TTS)

Issues in Serial Speech Translation

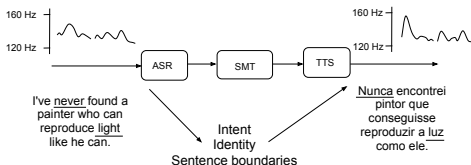
- A cascade of errors through the pipeline
 - Better integration needed (Mangu '07, Wolfel '08)
- Loss of information from the source side
 - ASR ignores source prosody, an essential part of speech

Need for tighter integration

Previous work

- Integration of ASR–SMT
 - Selection from multiple hypotheses
 - Translation of lattices
 - Al-Onaizan et al, '07, Wolfel et al, '08, Nöth '00
- Integration of SMT–TTS
 - Selection of translations optimal for synthesis
 - Adell et al., '12, Parlikar et al., '10
- Integration of ASR–TTS
 - Aguero et al, '06, Kurimo et al, '10

Our Goal



- Apply Intonation models and transformations to improve S2SMT (Motivated by Pisoni et al., '08)
- Impose source speaker characteristics on target side
 - Prominence patterns
 - Overall speaking style
 - Speaker identity

Parallel Speech Corpora

- A parallel speech corpus in English and Portuguese
- Airline magazine corpus TAP-UP used to select paragraphs to record
- A speaker fluent in both languages is recorded for both languages

Data statistics

Table: *Statistics of the EN-PT Parallel speech corpora*

	English	Portuguese
#Paragraphs	89	89
#Sentences	420	420
#Tokens	8184	8211
#Words	2934	3283
#Tokens/sentence	19.48	19.55
Duration(mins)	60.36	59.47

EMIME Parallel Speech Database

Table: The EMIME English-German parallel speech corpus

Language	English	German
Speaker ID	GM1	GM1
#Paragraphs	—	—
#Sentences	145	145
#Tokens	1301	1198
#Words	763	697
#Tokens/Sentence	8.97	8.26
Duration(in mins)	11.68	11.87

Manual analysis of Focus

- 75 sentences (10 mins) from the PT-EN corpus are manually annotated for Focus

Table: Results of manual annotation of focus in parallel speech

Language	Total #words	focussed words	non-focussed words	#focussed/ sentence
English	1569	298	1271	3.97
Portuguese	1585	285	1300	3.8

Manual Analysis ... contd

- 1100 word pairs aligned by GIZA++
 - 336 English words marked focussed
 - 303 English words marked focussed
- 48% word pairs have focus on both languages
- Comparable to inter listener disagreement within same language (Mo., '08)

Intent transfer in S2SMT (Anumanchipalli et al., 12)

- Transforming Intonation patterns over “comparable” content words
 - Fertility differences between the languages
 - Find word correspondence across the two languages

Experiment: Cross-lingual intent transfer

- Word level accents are parametrized under SPAM
 - Done for all content words of training data (90%)
 - word alignment is used to simulate word-level parallel data
- GMM-JDE Mapping function trained on the TILT parameter pairs

Intent transfer experiments

- SPSS synthetic voices are built for all databases
- Word level intonation models (Baseline)
- Source F0 transformed to estimate target F0

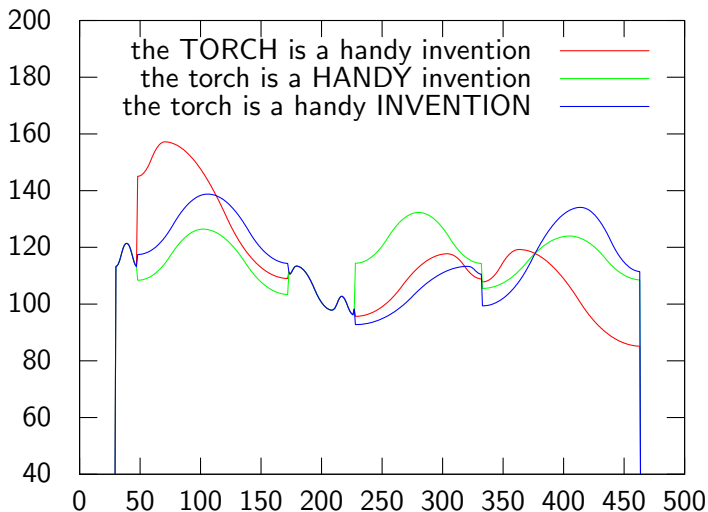
Table: Objective comparison of synthesized F0 contours

Lang Pair	Default		Proposed	
	rmse	corr	rmse	corr
en-pt	17.60	0.51	16.59	0.54
pt-en	15.90	0.47	15.30	0.49
en-de	11.93	0.54	10.98	0.51
de-en	10.27	0.46	10.17	0.46

Illustration of Prominence transfer

- A test set of 10 sentences is recorded with varying word focus
- Transformation applied on accent parameters of each content word

Focus Transfer



Example 1: 



Can this scale to Automatic dubbing ?

- Recorded videos from a Native Portuguese speaker
- Run through all stages of the S2SMT Pipeline

Demo

- Original Video
- Traditional ASR + SMT + TTS
- ASR + SMT + TTS/SPAM
- ASR + SMT + TTS/SPAM + Intent Transfer

Outstanding issues

- Preserving identity along with the style and intent across languages
- Synthesis sensitive to errors in ASR/SMT
- Efficient word and phone specific stretch/shrink functions
- Synthesizing paralinguistic events like laughter, hesitation

Summary

- A computational framework for prosodic description
 - Optimal modelling strategies to capture Intonation
 - Data-driven training and synthesis methods
 - Scalable synthesis across speakers, tasks and languages
- Voice conversion across speakers & languages
- Intent transfer for improved speech translation

Thank You!