

Segmental encoding of prosodic categories:

A perception study through speech synthesis

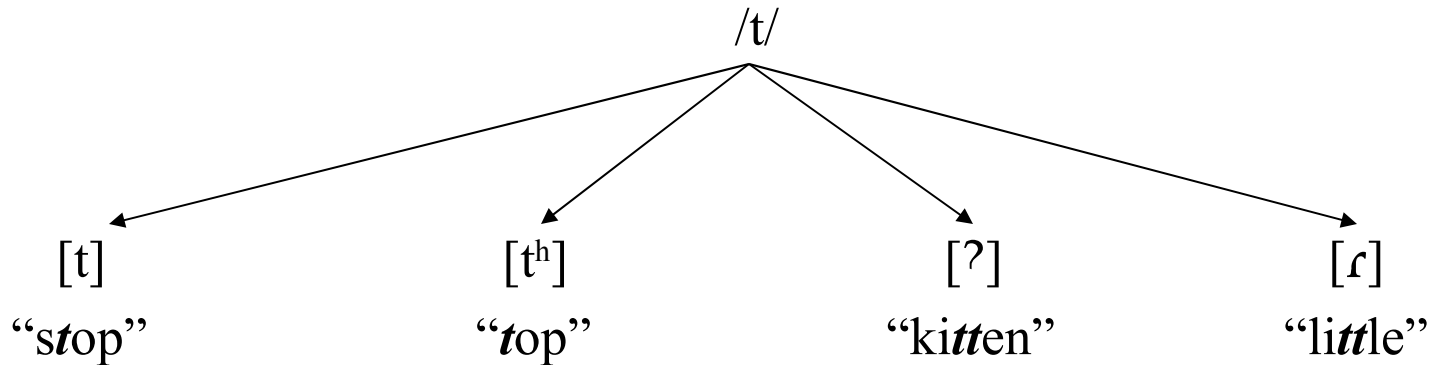


Kyuchul Yoon
2005. 8
The Ohio State University

Allophonic variations

- Defined mostly in terms of neighboring segments.

e.g. Allophones of /t/ in English

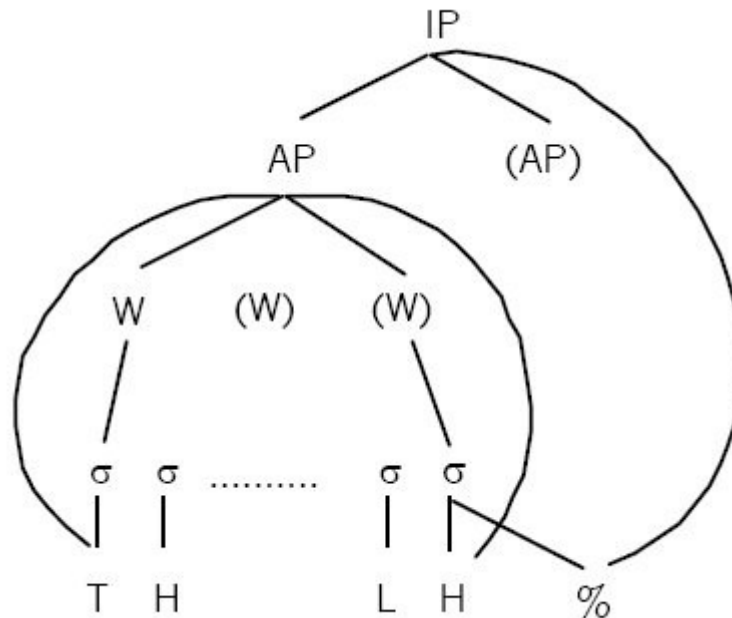


Segmental positions

- Determined in most cases within a word by its
 1. neighboring segments and
 2. word boundaries, i.e. word-initial/final
 3. presence/absence of stress

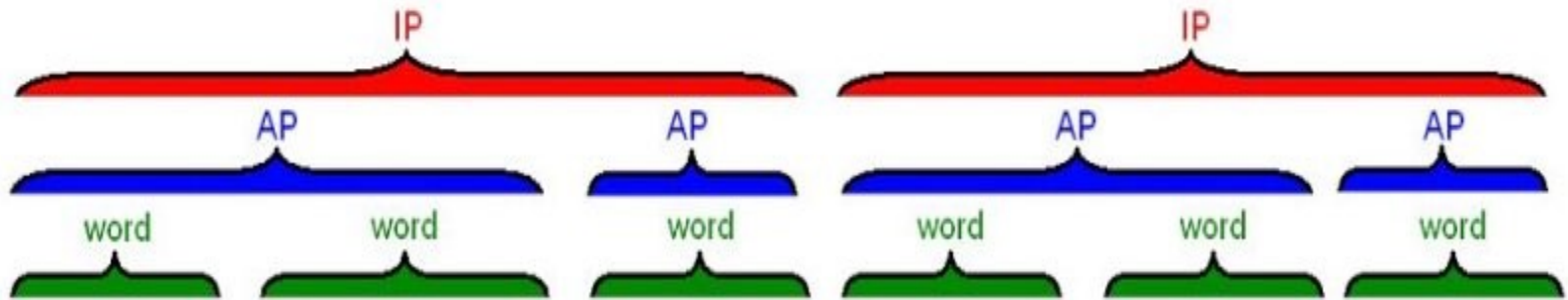
Korean Tone & Break Indices (K-ToBI)

(Prosody labeling conventions)

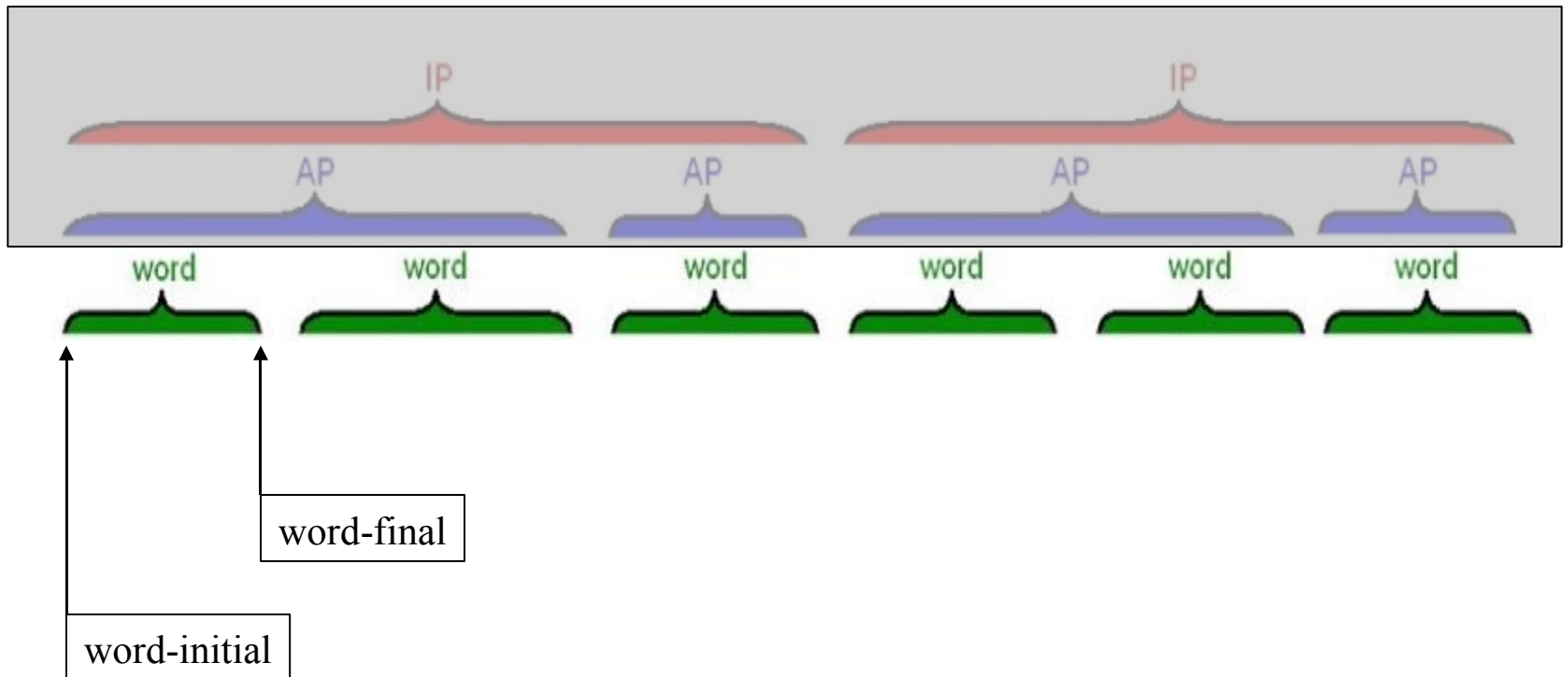


IP: Intonational Phrase	H: high tone
AP: Accentual Phrase	L: low tone
W: Prosodic Word (PW)	T: tone (could be H or L)
σ: syllable	°: boundary tone (e.g. H%, L%, HL%, etc.)

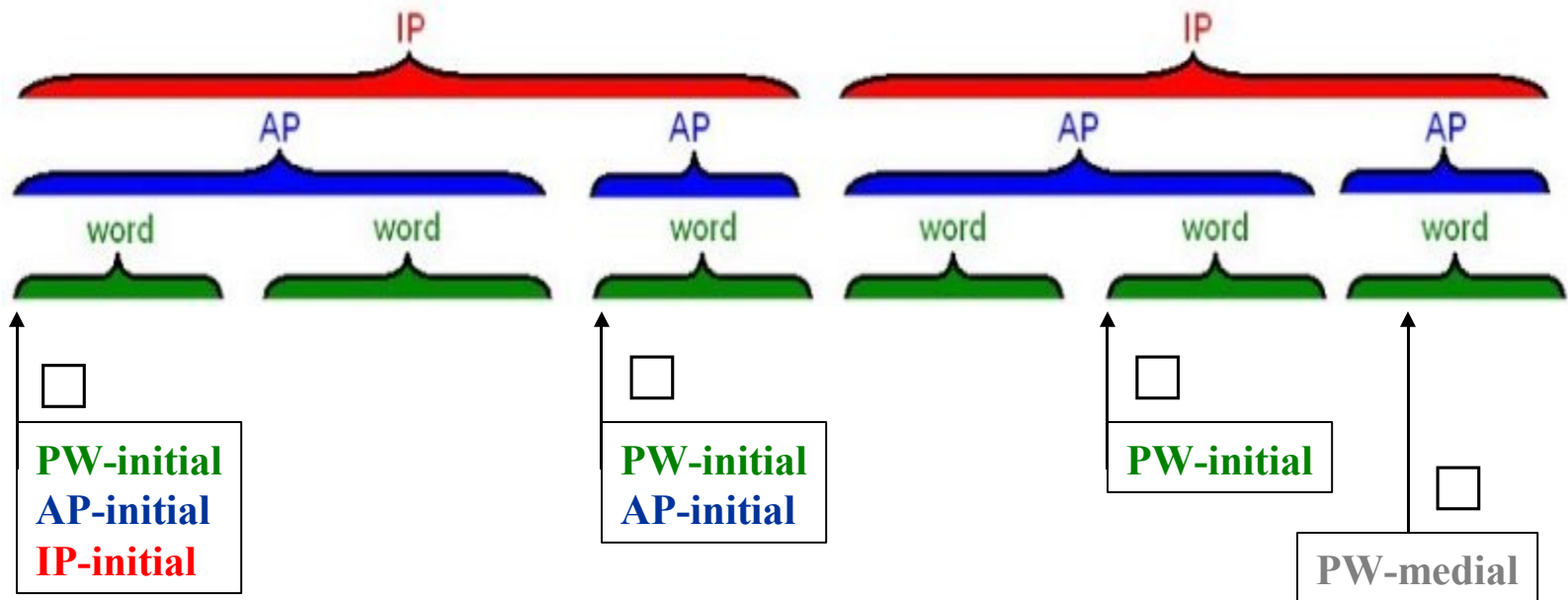
Word-initial positions in K-ToBI



Conventional segmental positions



Segmental positions in K-ToBI



Three types of word-initial positions in K-ToBI !

Allophonic variations:

an extended view

- Defined mostly in terms of neighboring segments.
- Need to be examined with respect to its prosodic constituency in K-ToBI.

Productions studies on Korean and other languages

- Korean

Jun ('93, '98): lenis stop voicing, obstruent nasalization, VOT of /p^h/

Cho & Keating ('01): segmental properties of /t, t^h, t*, n/

Kim ('01): segmental properties of /s^h, s*/

Yoon ('03): subsegmental durations of /s^h, s*/

- Other languages

Smith ('97): American /z/

Pierrehumbert & Talkin ('92), Pierrehumbert ('95): English /h/ and /ʔ/

Fougeron ('01): French segments /t, k, s, l, n, i, a/

Keating et al. ('98): /t, n/ of Korean, English, French & Taiwanese

Productions studies

on Korean and other languages – summary of results

- Korean

AP is the domain of lenis stop voicing, post-obstruent tensing (Jun).

IP is the domain of obstruent nasalization (Jun).

VOT of /p^h/: AP-initial > PW-initial > PW-medial (Jun).

Consonants initial to higher prosodic domains are ‘stronger’ (Cho, Keating, Kim).

Non-uniform variations in durations of subsegmental units (Yoon).

- Other languages

American English /z/ is devoiced differently in different positions (Smith).

English /h/ and /ʔ/ produced differently in different word-/phrase-level prosody. (P & T)

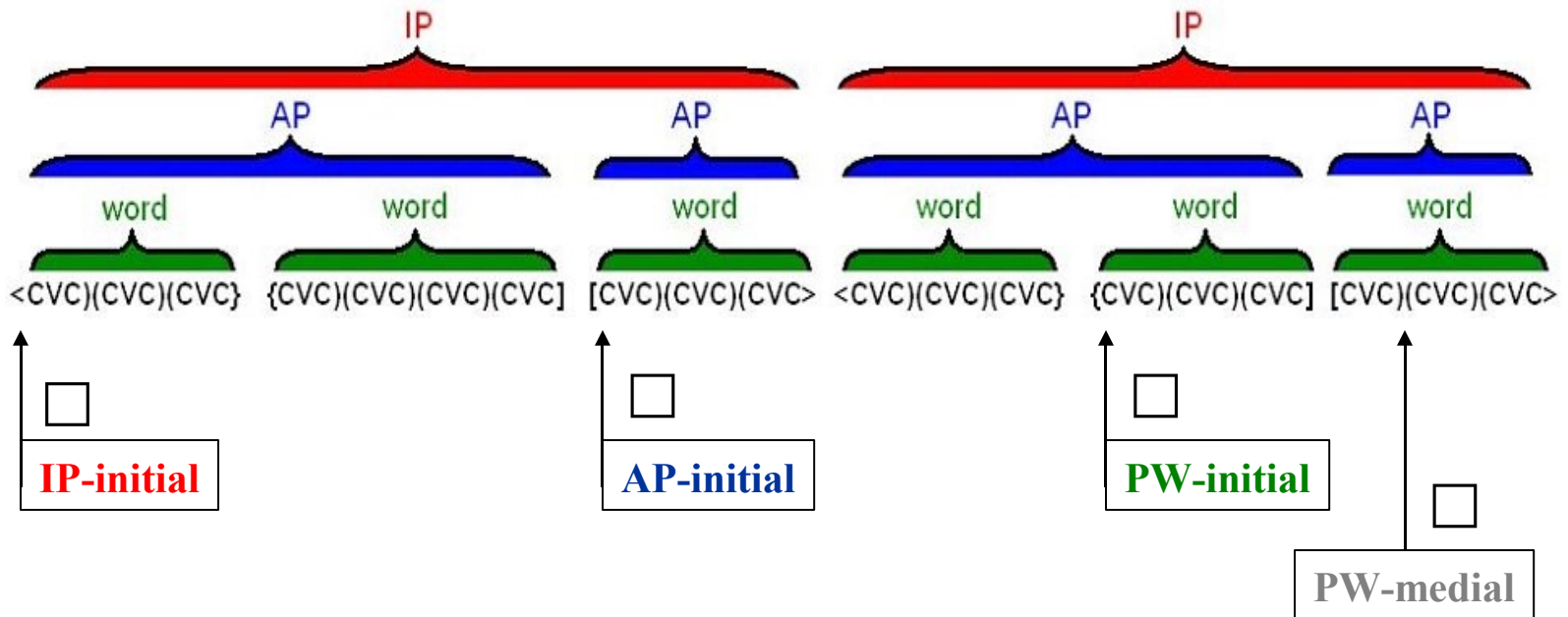
Articulation of initial segments varied depending on the prosodic level of the constituent, i.e. initial to an IP, AP, W or syllable. (Fougeron)

There is phrasal/prosodic conditioning of articulation across the four languages.
(Keating et al.)

Need for a perception study, but how?

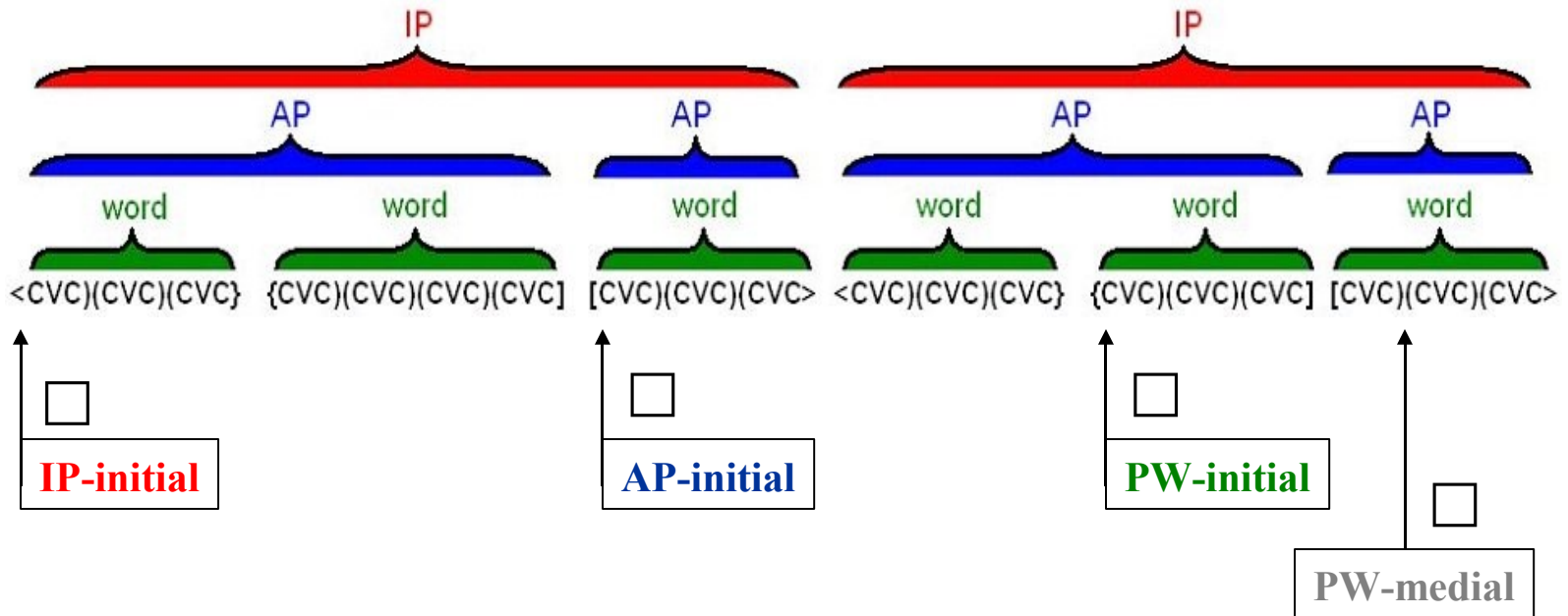
- As the production studies show, Korean speakers seem to *encode* prosodic categories, i.e. IP, AP, PW, etc., in domain-initial segments.
- Do speakers *decode* the encodings?
Are the encodings perceptible?
- How do we test it?
One way to test it is to use a concatenative TTS system so that one can synthesize sentences by manipulating phone-sized units, i.e. diphones. (*Festival Speech Synthesis System*)

Need for a perception study, but how?



Key idea: Synthesize a set of two sentences, differing only in terms of their domain-initial segment compositions.

Need for a perception study, but how?



Test stimuli:

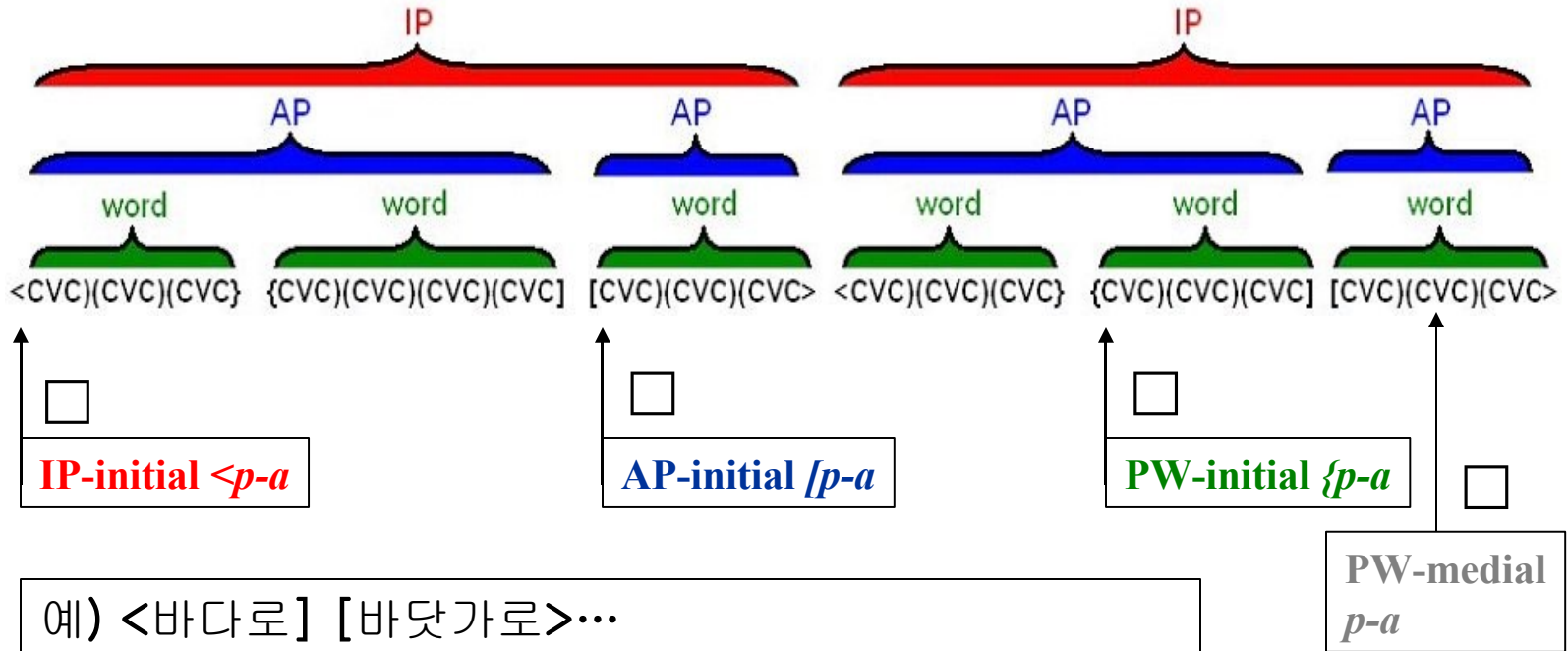
1st set: **good AP**: composed of prosodically appropriate synthetic units

bad AP: composed of prosodically *inappropriate* units (Replace ② with ④)

2nd set: **good PW**: composed of prosodically appropriate synthetic units

bad PW: composed of prosodically *inappropriate* units (Replace ③ with ④)

Prosodic diphones



6,503 prosodic diphones needed to synthesize *any* Korean utterance.

Design & synthesis of test stimuli

- 96 stimuli (phrases) synthesized from the Festival system (Durations and F0 contours copied from natural utterances).
- All were composed of either two AP's or two PW's.
- All contained one target site, where an AP/PW-initial segment was replaced with a PW-medial segment.

24 *good AP*: phrases with intact diphones.

24 *bad AP* : phrases whose target site segment (AP-initial segment) was replaced with a PW-medial segment

24 *good PW*: phrases with intact diphones

24 *bad PW* : phrases whose target site segment (PW-initial segment) was replaced with a PW-medial segment

Design & synthesis of test stimuli

- Prototype system lacks duration & F0 generation module
⇒ Get help from natural utterances.

- Synthesis of a sample stimulus (*Praat script*)

<삼성 차의] [가치는>



natural utterance



diphone sequences from Festival



fundamental frequency (F0) contour and segmental durations
copied from natural utterance

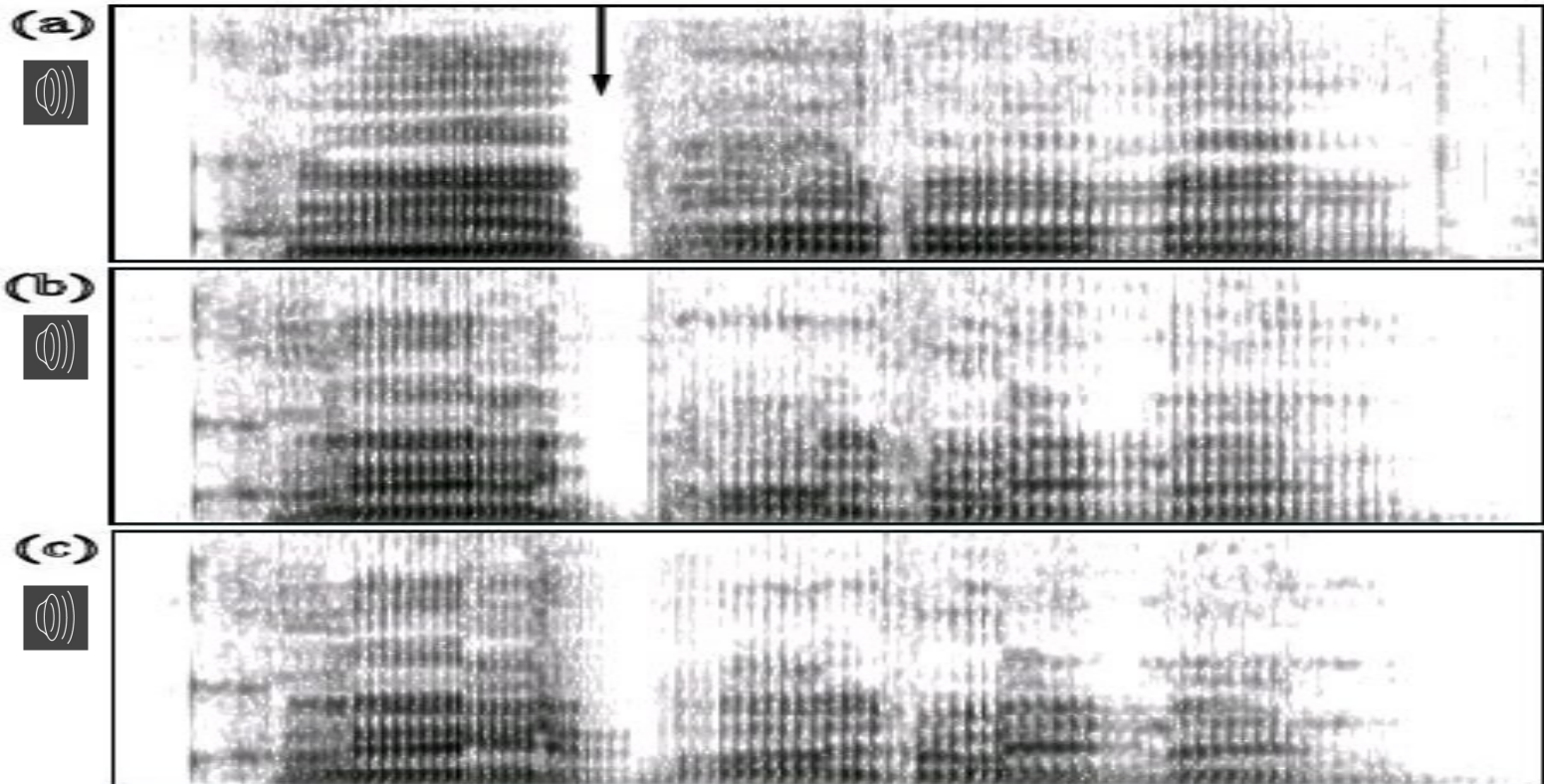


intensity contour copied from natural utterance

Design & synthesis of test stimuli





































- Sample stimuli

<그의] [발언은> target site segment: /p/



Design & synthesis of test stimuli

- More sample stimuli

target segment	good AP	bad AP		good PW	bad PW
/p/					
/t/					
/k/					
/p ^h /					
/t ^h /					
/t [*] /					
/tʃ/					
/tʃ ^h /					
/s ^h /					

Results & conclusion

- 80 listeners (37 women and 43 men):
native speakers of Korean, average age of 30.6, grew up in Korea until at least 18 years old.
- Two types of tests in three tasks
Intelligibility: dictation task
⇒ wrote down what they heard in hangul
Naturalness: rating & preference task
⇒ rate one version wrt/ the other and
⇒ choose one over the other
- Three factor ANOVAs
Factor I: appropriateness (“good” vs. “bad”)
Factor II: break level (AP vs. PW)
Factor III: consonant type (lenis vs. non-lenis)

Results & conclusion

	By items			
	play count	# of <i>eojeol</i> correct	rating	choice
Factor I (appropriateness)	F(1,88)=7.50 p<0.01		F(1,88)=36.34 p<0.001	F(1,88)=50.22 p<0.001
Factor II (break level)				
Factor III (consonant type)				
Interaction			F(1,88)=15.11 p<0.001 (I×III)	F(1,88)=5.13 p<0.05 (I×II) F(1,88)=5.92 p<0.05 (I×III)

	By subjects			
	play count	# of <i>eojeol</i> correct	rating	choice
Factor I (appropriateness)	F(1,632)=22.57 p<0.001	F(1,632)=16.15 p<0.001	F(1,632)=44.11 p<0.001	F(1,632)=139.72 p<0.001
Factor II (break level)		F(1,632)=20.01 p<0.001		
Factor III (consonant type)				
Interaction			F(1,632)=5.06 p<0.05 (I×II) F(1,632)=27.34 p<0.001 (I×III)	F(1,632)=24.65 p<0.001 (I×II) F(1,632)=14.28 p<0.001 (I×III)

Results & conclusion

- Statistical analyses showed that listeners performed better in the dictation task with “good” versions of the stimuli. They also liked/rated better the “good” versions.
- **Segmental encoding of prosodic domains/categories seems perceptible to Korean listeners.**