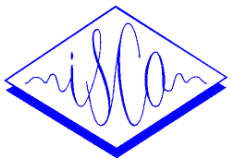


First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

SYLLABLE STRUCTURE PARSING FOR CONTINUOUS SPEECH RECOGNITION

Shigeru Ono

C & C Information Technology Research Laboratories
NEC Corporation
4-1-1, Miyazaki, Miyamae-ku, Kawasaki 213, Japan

Abstract

This paper describes a scheme to deal with allophonic and coarticulatory variations for phoneme-based continuous speech recognition and a probabilistic algorithm to parse syllable structure from acoustic speech realizations. In the scheme, phonological objects are represented in terms of *"syllable features"*–syllable positions– and *"phoneme features"* –distinctive features–, and they are organized within hierarchical structures. The constituent features of the structures are associated with the acoustic realizations through probabilistic measure. In the algorithm, syllable structure is parsed from the acoustic realizations by applying the acoustic-phonological constraints and the collocational restrictions involved in the internal constituent features. Performance results for 15 test sentences spoken by 5 male speakers that phonemes are recognized at 90.5% accuracy, and syllable structure is parsed at 79.7% accuracy.

1 Introduction

For phoneme-based speech recognition, allophonic and coarticulatory variations have been considered to be problematic, or a kind of noise that makes it more difficult to hypothesize lexical candidates. On the other hand, an alternative view, wherein allophonic variations are essential to reveal important characteristics of supersegmental context, has been recently offered. In particular, Church[1] and Randloph[9] suggested and argued, respectively, that acoustic speech realization variations would be predictable and could be systematically dealt within an augmented context-free grammar, if they are restricted to those which are caused by syllable structure. They also demonstrated how the syllable structure can be parsed and how recognition can be conducted by using the parsed structure, if detail allophonic, or acoustic, representations are given. Moreover, the relevant works from the perceptual viewpoint [6][8] also showed that allophonic variations serve as prime markers of syllable, word, and larger morpho-syntactic boundaries. These reports are attractive and motivating for continuous speech recognition application, but their approach is not directly applicable to actual problems, because it is quite difficult to extract the "detailed" allophonic, or acoustic, representations from the acoustic speech realizations.

This paper presents a scheme based on phonology to deal with allophonic and coarticulatory variations and a concrete probabilistic algorithm to parse syllable structure form the acoustic realizations. In the scheme, allophonic representation is avoided, wherein phonological objects are defined in terms of *"syllable features"*–syllable positions– and *"phoneme features"*–distinctive features–; allophonic and coarticulatory variations are considered as being caused by the conditions of the feature collocations. The features are associated with the acoustic realizations through probabilistic measure. In the algorithm, the syllable structure is parsed from the acoustic speech realizations, by using the acoustic-phonological constraints and the restrictions on the feature collocations.

In the following, the definition of the phonological object discussed in this paper will be presented, and then the parsing algorithm will be given. Finally, experimental results for continuous speech will be shown.

2 Definition of Phonological Object

In this paper, phonological object is defined as syllable concatenation.

$$[\ PHONOLOGY \quad [SYLLABLE \cdots SYLLABLE]\ ]$$

The syllable structure[9] is defined as shown in Table 1, where the terminal constituent values are concatenation of phonemes.

$$\left[\begin{array}{l} SYLLABLE: < PHONEME \cdots PHONEME > \\ \left[\begin{array}{lll} & ONSET & \left[\begin{array}{l} OUTER-ONSET \\ INNER-ONSET \end{array}\right] \\ CORE & & \\ & RHYME & \left[\begin{array}{ll} NUCLEUS & \\ CODA & \left[\begin{array}{l} INNER-CODA \\ OUTER-CODA \end{array}\right] \end{array}\right] \\ AFFIX & [\ AFFIX-i\ ] & \end{array}\right] \end{array}\right]$$

Table 1: Syllable Structure

The sonority sequencing principles prescribe a specific assignment of phonemes to the terminal constituents as follows:

OUTER-ONSET	→	Stop | Strong Fricative | Weak Fricative | Affricative
INNER-ONSET	→	Nasal | Semivowel
NUCLEUS	→	Vowel
INNER-CODA	→	Nasal | Semivowel
OUTER-CODA	→	Stop | Strong Fricative | Weak Fricative | Affricate | Nasal | Semivowel
AFFIX-i	→	Stop | Fricative

The phoneme structure [10] is defined in terms of the distinctive features as depicted in Table 2, where the terminal constituents values are '+' or '-'.

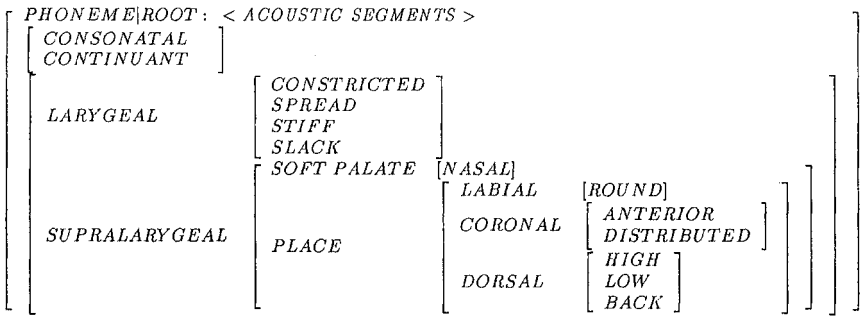


Table 2: Phoneme Structure

Thus, the above phonological object is well defined by using the features within the hierarchical structures. The advantages of these types of representations, based on features and hierarchical structures, have been well investigated, for example, by [2][4][9][10][12].

One of the favorable properties of these representation is the computational efficiency in filtering out undesired hypotheses and identifying correct entries from lexical candidates. Randloph[9] argued that the syllable structure defined above is described in a context-free grammar, and that the hierarchical structure reflects the magnitude of collocational constraints within syllable. Moreover, Clements[2] and Sagey[10] claimed that the phonological variations –assimilation, gemination, or deletion– are formalized as manipulations of the partial feature group of the phoneme structure, which means that the attributes of the adjacent phonological segments can be predicted by manipulating the group of features of the current phonological segment.

3 Phonological Parsing Algorithm

The mapping relation between the acoustic realizations and the phonological objects are not one-to-one. Therefore, in order to define the relation mathematicaly, probability measure is introduced in this paper. Adopting probability measure, the mapping function from the acoustic realizations to the phonological objects is written as

$$P(\Lambda|\Theta, A)P(\Theta|A) \tag{1}$$

where Λ is a set of the syllable concatenation, Θ is a set of the phoneme concatenation, and A is a set of the acoustic segment concatenation.

Using this probabilistic representation, the phonotactic constraints, the phonotactic constraints in the syllable structure, and the acoustic-phonological constraints are represented as $P(\Theta)$, $P(\Lambda|\Theta)$, and $P(\Lambda, \Theta|A)$, respectively.

To calculate Equation 1 for all possible combination among Λ, Θ, and A, the huge number of training data is required. So, in this paper the following four assumption are adopted.

ASSUMPTION:

- The acoustic speech realizations among the acoustic segments are mutually independent.
- The phoneme features are related to each other only in three successive phonological segments.
- The phoneme features of the left, or right, phonological segment are not uniquely related to the current acoustic segment. The acoustic realization of the acoustic segment is influenced by a partial group of distinctive features of the adjacent phonological segments.

Incorporating these hypotheses into Equation 1, it is written as follows:

$$\begin{aligned} & P(\Lambda|\Theta, A)P(\Theta|A) \\ &= \prod_i P(\lambda_i, \theta_i|\lambda_{i-1}, \theta_{i-1}, \lambda_{i-2}, \lambda_{i-2}) \\ & \quad \prod_i P(g(\theta_i), \theta_{i-1}, \lambda_{i-1}, g(\theta_{i-2})|\lambda_i, \theta_i, \lambda_{i-1}, \theta_{i-1}, \lambda_{i-2}, \theta_{i-2}) \\ & \quad \prod_j \frac{P(g(\theta_j{}_{-1}), \theta_j, \lambda_j, g(\theta_{j+1})|a_i)}{P(g(\theta_j{}_{-1}), \theta_j, \lambda_j, g(\theta_{j+1}))} \end{aligned} \tag{2}$$

where $g()$ represents a group of the distinctive features.

Therefore, to extract the phonological feature values from the acoustic realizations is equivalent to detect the Λ and the Θ maximizing Equation 2 for the given A.

$$F(\Lambda, \Theta) = \max_{\Lambda, \Theta} P(\Lambda, \Theta|A) \tag{3}$$

In calculating Equation 3, It should be noted that the number of acoustic segment is not equal to that of phonological segment; in assimilation, gemination, or deletion segment, the number of acoustic segment is less than that of phonological segment by one. The relationship between acoustic segment and phonological segment for allophonic variations is illustrated in Fig. 1.

Therefore, an algorithm to calculate Equation 3 is as follows:

ALGORITHM:

The following iterations are carried out for n from 2 to $2M+1$, in which M is the number of the acoustic segments.

$$\begin{aligned} & F_{1(n,m)}(\lambda_k, \theta_k|\lambda_{max}, \theta_{max}, F_{max}) \\ &= \max_{F_{1(n-1,m-1)}, F_{3(n-1,m-1)}} \\ & \quad \{\max_{\lambda_i, \theta_j} \{F_{1(n-1,m-1)}(\lambda_i, \theta_i|\lambda_{max}, \theta_{max}, F_{max}) \\ & \quad + \log P(\lambda_k, \theta_k|\lambda_{i-1}, \theta_{i-1}, \lambda_{max}, \theta_{max}) \\ & \quad + \log P(g(\theta_k), \theta_i, \lambda_i, g(\theta_{max})|\lambda_k, \theta_k\lambda_{i-1}, \theta_{i-1}, \lambda_{max}, \theta_{max}) \\ & \quad + \log P(g(\theta_k), \theta_i, \lambda_i, g(\theta_{max})|a_m) - \log P(g(\theta_k), \theta_i, \lambda_i, g(\theta_{max})), \\ & \quad F_{3(n-1,m-1)}(\lambda_i, \theta_i|\lambda_{max}, \theta_{max}, F_{max}) + \log P(\lambda_k, \theta_k|\lambda_{i-1}, \theta_{i-1}, \lambda_{max}, \theta_{max}) \end{aligned}$$

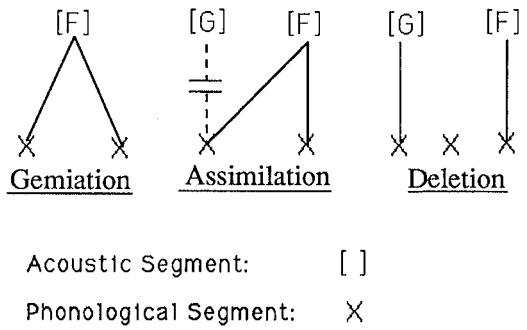


Figure 1: Relationship between acoustic segments and phonological for allophonic variations

$$
\begin{aligned}
&+\log P(g(\theta_k),\theta_i,\lambda_i,g(\theta_{max})|\lambda_k,\theta_k\lambda_{i-1},\theta_{i-1},\lambda_{max},\theta_{max})\\
&+\log P(g(\theta_k),\theta_i,\lambda_i,g(\theta_{max})|a_m)-\log P(g(\theta_k),\theta_i,\lambda_i,g(\theta_{max}))\}
\end{aligned}
$$

$$
\begin{aligned}
&F_{2(n,m)}(\lambda_k^d,\theta_k^d|\lambda_{max},\theta_{max},F_{max})\\
&=\max_{F_{1(n-1,m)},F_{3(n-1,m)}}\\
&\{\max_{\lambda_i,\theta_j}\{F_{1(n-1,m)}(\lambda_i,\theta_i|\lambda_{max},\theta_{max},F_{max})\\
&+\log P(\lambda_k^d,\theta_k^d|\lambda_{i-1},\theta_{i-1},\lambda_{max},\theta_{max})\\
&+\log P(g(\theta_k^d),\theta_i,\lambda_i,g(\theta_{max})|\lambda_k^d,\theta_k^d\lambda_{i-1},\theta_{i-1},\lambda_{max},\theta_{max})\\
&+\log P(g(\theta_k^d),\theta_i,\lambda_i,g(\theta_{max})|a_m)-\log P(g(\theta_k^d),\theta_i,\lambda_i,g(\theta_{max})),\\
&F_{3(n-1,m)}(\lambda_i,\theta_i|\lambda_{max},\theta_{max},F_{max})+\log P(\lambda_k^d,\theta_k^d|\lambda_{i-1},\theta_{i-1},\lambda_{max},\theta_{max})\\
&+\log P(g(\theta_k^d),\theta_i,\lambda_i,g(\theta_{max})|\lambda_k^d,\theta_k^d\lambda_{i-1},\theta_{i-1},\lambda_{max},\theta_{max})\\
&+\log P(g(\theta_k^d),\theta_i,\lambda_i,g(\theta_{max})|a^d)-\log P(g(\theta_k^d),\theta_i,\lambda_i,g(\theta_{max}))\}
\end{aligned}
$$

$$
\begin{aligned}
&F_{2(n,m)}(\lambda_k,\theta_k|\lambda_{max},\theta_{max},F_{max})\\
&=\max_{F_{1(n-1,m)},F_{3(n-1,m)}}\\
&\max_{\lambda_i,\theta_j}\{F_1(n-1,m)(\lambda_i,\theta_i|\lambda_{max},\theta_{max},F_{max})\\
&+\log P(\lambda_k,\theta_k|\lambda_{i-1},\theta_{i-1},\lambda_{max},\theta_{max})\\
&+\log P(g(\theta_k),\theta_i,\lambda_i,g(\theta_{max})|\lambda_k,\theta_k\lambda_{i-1},\theta_{i-1},\lambda_{max},\theta_{max})\\
&+\log P(g(\theta_k),\theta_i,\lambda_i,g(\theta_{max})|a_m)-\log P(g(\theta_k),\theta_i,\lambda_i,g(\theta_{max})),\\
&F_{3(n-1,m)}(\lambda_i,\theta_i|\lambda_{max},\theta_{max},F_{max})+\log P(\lambda_k,\theta_k|\lambda_{i-1},\theta_{i-1},\lambda_{max},\theta_{max})\\
&+\log P(g(\theta_k),\theta_i,\lambda_i,g(\theta_{max})|\lambda_k,\theta_k^d\lambda_{i-1},\theta_{i-1},\lambda_{max},\theta_{max})\\
&+\log P(g(\theta_k),\theta_i,\lambda_i,g(\theta_{max})|a_m)-\log P(g(\theta_k),\theta_i,\lambda_i,g(\theta_{max}))\}
\end{aligned}
$$

$$
\begin{aligned}
&F_{3(n,m)}(\lambda_k,\theta_k|\lambda_{max},\theta_{max},F_{max}=F_2)\\
&=\max_{\lambda_i=\lambda_i^d\ or\ \lambda_i,\theta_i=\lambda_i^d\ or\ \lambda_i}\{F_{2(n-1,m-1)}(\lambda_i,\theta_i|\lambda_{max},\theta_{max},F_{max})\\
&+\log P(\lambda_k,\theta_k|\lambda_{i-1},\theta_{i-1},\lambda_{max},\theta_{max})\\
&+\log P(g(\theta_k),\theta_i,\lambda_i,g(\theta_{max})|\lambda_k,\theta_k\lambda_{i-1},\theta_{i-1},\lambda_{max},\theta_{max})\\
&+\log P(g(\theta_k),\theta_i,\lambda_i,g(\theta_{max})|a_m)-\log P(g(\theta_k),\theta_i,\lambda_i,g(\theta_{max}))\}
\end{aligned}
$$

where λ_{max} and θ_{max} correspond to the λ_i and the θ_i maximizing F_i, respectively, and F_{max} for F_1 and F_3 is F_1 or F_3. λ_k^d and θ_k^d are syllable position and distinctive features selected at the deleted acoustic segments, respectively.

The group of the distinctive features used in this paper is listed in Table 3. Phonemes are redundantly assigned to appropriate groups. For example, phoneme /o/ or /u/ is assigned to ":DORSAL+HIGH-LOW+BACK" and ":LABIAL+CONT-CONS" . This list indicates that the phonemes assigned to the same distinctive feature groups should provide the same acoustical effects on the adjacent phonemes.

$: LABIAL - CONT + CONS$
$: LABIAL + CONT - CONS$
$: LABIAL + CONT + CONS$
$: CORONAL - CONT - CONS$
$: CORONAL + CONT - CONS$
$: CORONAL + CONT + CONS$
$: DORSAL - CONT + CONS$
$: DORSAL - HIGH - LOW - BACK$
$: DORSAL - HIGH - LOW + BACK$
$: DORSAL - HIGH + LOW - BACK$
$: DORSAL - HIGH + LOW + BACK$
$: DORSAL + HIGH - LOW - BACK$
$: DORSAL + HIGH - LOW + BACK$
$: DORSAL + CONT + CONS$
$: LABIAL - NASL$
$: CORONAL - NASAL$
$: DORSAL - NASAL$

Table 3: List of the distinctive feature groups

	Training data	*Test data*
Phoneme Recognition	94.6 %	90.5 %
Syllable Parsing	86.4 %	79.7 %

Table 4: Phoneme Recognition Accuracy and Syllable Parsing Accuracy

4 Experiments

4.1 Data Description

The source data for the experiments is a subset of the TIMIT acoustic phonetic database [6]. The utterance data is manually segmented into phonemic units. 600 sentences spoken by 120 male speakers are used as training samples to estimate the probability function shown in Equation 2. 15 unknown sentences spoken by 5 male speakers are used as test samples to evaluate the proposed parsing algorithm.

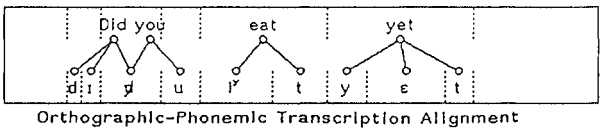
4.2 Acoustic Speech Representation

In the experiment, the acoustic speech signals were transformed to 40 channel Seneff's hair-cell envelopes every 5 msec. The envelopes and the differentiated envelopes for adjacent frames were averaged over segment; Glass[3] claimed that the averaged envelopes has good resolution to capture consistent contextual dependencies. The two kinds of the envelopes were independently quantized at 9 bits using vector quantization techniques[7].

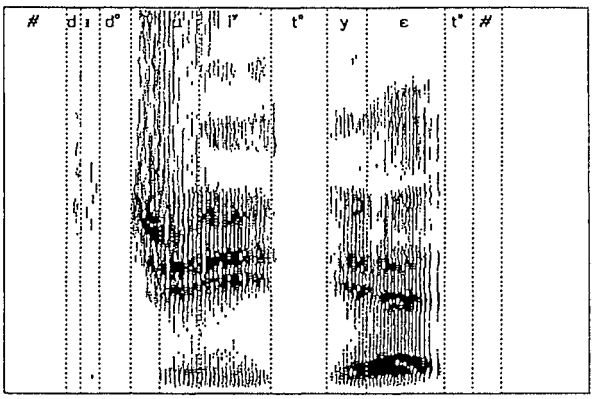
4.3 Performance Results

The performance results are summarized in Table 1. The results for the training data were obtained from a subset of the training database: 122 sentences spoken by 25 male speakers.

An example of parsed results is illustrated in Fig. 3. In this example, the assimilation between /d/ and /j/ and the deletion of /t/ were successfully parsed. In addition, all the phonemes and the syllable structure were correctly recognized.



Orthographic-Phonemic Transcription Alignment



Wide-Band Spectrogram

```
Reference Syllable Structure:
(((1 'd' :OUTER-ONSET) (2 'I' :NUCLEUS) (3 'd' :CODA-AFFIX))
 ((4 'y' :INNER-ONSET) (5 'u' :NUCLEUS))
 ((6 't' :NUCLEUS) (7 't' :CODA-AFFIX))
 ((8 'y' :INNER-ONSET) (9 'E' :NUCLEUS) (10 't' :CODA-AFFIX)))

Parsed Syllable Structure:
(((1 'd' :OUTER-ONSET) (2 'I' :NUCLEUS) (3 'd' :CODA-AFFIX))
 ((4 'y' :INNER-ONSET) (5 'u' :NUCLEUS))
 ((6 't' :NUCLEUS) (7 't' :CODA-AFFIX))
 ((8 'y' :INNER-ONSET) (9 'E' :NUCLEUS) (10 't' :CODA-AFFIX)))
```

Figure 2: Example of parsed results

5 Conclusion

A scheme based on phonological theories for dealing with allophonic and coarticulatory variations for continuous speech recognition and a syllable parsing algorithm to construct syllable structure from acoustic speech realizations were presented. In the scheme, the phonological objects were represented in terms of *"syllable features"*–syllable positions– and *"phoneme features"* –distinctive features–, and they were organized within hierarchical structures. The constituent features of the structures were associated with the acoustic realizations through probabilistic measure. In the algorithm, the syllable structure was parsed from the acoustic realizations by applying the acoustic-phonological constraints and the collocational restrictions on the internal constituent features.

The experiments for 15 test sentences spoken by 5 male speakers showed that phoneme were recognized at 90.5 % accuracy and syllable structure was parsed at 79.7 % accuracy. This experimental results indicate great promise for this scheme for speech recognition application.

Acknowledgements

This work was conducted while the author was a visiting scientist at the Research Laboratory of Electronics and the Laboratory for Computer Science at MIT. The author would like to acknowledge the encouragement and the support of Dr. Victor Zue and Prof. Ken Stevenes. He would also like to thank members of the MIT Spoken Language Systems group, particularly Jim Glass, Hong Leung, Mike Phillips, John Pitrelli, and Mark Randloph, for their usefull suggestions and software support.

References

[1] Church, K. W., "Phrase Structure Parsing: A Method for Taking Advantage of Allophonic Constraints," Ph. D. dissertation, Massachusetts Institute of Techonlogy, MA, 1983.

[2] Clements, G. N., "The Geometry of Phonological Features," Phonology, No. 2, pp.225-252, 1985.

[3] Glass, J. R., "Finding Acoutic Regularities in Speech: Applications to Phonetic Recognition," Ph. D. dissertation, Massachusetts Institue of Technology, Cambridge, MA, 1988.

[4] Kahn, D.,"Syllable-based Generalizations in English Phonology, Ph. D. dissertation, Massachusetts Institute of Techonlogy, MA, 1976.

[5] Lamel, L. F., Kassel, R. H., and Seneff.S., "Speech Database Development:Design and Analysis of the Acoustic-phonetic Corpus," *Proc. DARPA Speech Recognition Workshop*, Report No. SAIC-86/1546, pp.100-109, 1986.

[6] Lamel, L. F., "Formalizing Knowledge used in Spectrogram Reading: Acoustic and perceptual evidence from stops," Ph. D. dissertation, Massachusetts Institute of Technology, MA, 1988.

[7] Makhol, J., Roucos, S., and Gish, H.,"Vector Quantization in Speech Coding," *Proc. IEEE*, Vol.73, No. 11, pp.1551-1588, 1985.

[8] Nakatani, L. and Dukes, K.,"Locus of Segmental Cues for Word Juncture," *J. Acoust. Soc. Amer.*, Vol. 62, No. 3, pp.714-719, 1977.

[9] Randolph, M. A., "Syllable-based Constraints on Properties of English Sounds," Ph. D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1989.

[10] Sagey, E. C., "The Representation of Features and Relations in Non-linear Phonology," Ph.D. dissertation, Massachusetts Institutes of Technology, MA, 1986.

[11] Seneff, S.,"A joint synchrony/mean-rate model of auditory speech processing," *J. Phonetics*, Vol. 16, No. 1, pp.55-76, 1988.

[12] Stevens, K. N.,"Phonetic Features and Lexical Access," unpublished report, 1989.