Vowel Recognition from Articulatory Position Time-Series Data

Jun Wang[1], Ashok Samal[1], Jordan R. Green[2], Tom D. Carrell[2]
[1]Department of Computer Science & Engineering, {junwang, samal}@cse.unl.edu
[2]Department of Special Education & Communication Disorders, {jgreen4, tcarrell1}@unl.edu
University of Nebraska-Lincoln, Lincoln, NE, USA

***Abstract*--A new approach of recognizing vowels from articulatory position time-series data was proposed and tested in this paper. This approach directly mapped articulatory position time-series data to vowels without extracting articulatory features such as mouth opening. The input time-series data were time-normalized and sampled to fixed-width vectors of articulatory positions. Three commonly used classifiers, Neural Network, Support Vector Machine and Decision Tree were used and their performances were compared on the vectors. A single speaker dataset of eight major English vowels acquired using Electromagnetic Articulograph (EMA) AG500 was used. Recognition rate using cross validation ranged from 76.07% to 91.32% for the three classifiers. In addition, the trained decision trees were consistent with articulatory features commonly used to descriptively distinguish vowels in classical phonetics. The findings are intended to improve the accuracy and response time of a real-time articulatory-to-acoustics synthesizer.**

***Keywords - articulatory speech recognition; time-series; neural network; support vector machine; decision tree*.**

I. INTRODUCTION

Recognizing speech from articulatory movements has attracted the attention of researchers from a variety of fields (i.e., computer science, electronic engineering, and communication science and disorders) in recent years. Potential applications include a robust objective technique for assessing the severity of speech impairment and facilitating the treatment of speech disorders. For example, a real-time articulatory-to-acoustic synthesizer based on this technique may enable laryngectomee patients to speak in a more natural way than is currently possible. It can also be used as a tool for speech motor learning research [17] and to facilitate second language learning [14]. In addition, a better understanding of the relations between articulatory movements and speech will facilitate acoustic speech recognition [5, 22], and the development of text-to-speech synthesis (TTS) that are based on articulatory modeling.

However, the complexity of human speech production mechanism makes articulatory speech recognition a formidable problem. One significant challenge is the many-to-one mapping between articulatory configurations and phonemes. The mapping has not been clearly determined. In classical phonetics, phonemes are distinguished by their location of primary constriction in the vocal tract (i.e., dental, velar) and the manner (i.e., stop, continuant) in which they are produced. However, attempts to develop mappings between speech movements and their associated sounds have been significantly challenged by the high degree of variation within and across talkers in their lip, jaw, and tongue movements.

There are mainly two types of approaches to recognize speech from articulatory movements in literature. The first type is to extract articulatory features (i.e., maximum mouth width/height, tongue dorsum peak position, maximum mouth opening/closing velocity) from articulatory position data, then to recognize phonemes from those features [3, 4, 5, 6, 7]. Blackburn and Young [8] trained a pseudo articulatory speech synthesis model using x-ray data of speech movement for recognition. This type of approaches is based on the phonetic knowledge for distinguishing vowels mentioned in the previous paragraph. The second type of approaches is to map articulatory position data (including facial motion) to acoustic data [9, 10, 11, 12, 13]. High correlations have been found between articulatory movement data and acoustics of syllables [10] or sentences [12, 13]. Classifiers such as Neural Network and Hidden Markov Models were used in both approaches [4, 5, 6, 12]. Although some of the results are encouraging, articulatory speech recognition research is still in its early stages. The recognition rates obtained in the literature ranged from 70s to low 90s in percentage, and most of the results have been obtained from relatively small datasets.

Our long-term goal is to build a real-time articulatory-to-acoustics synthesizer. The device will take articulatory position time-series data as input, recognize the speech (phonemes, syllables or words), then synthesize a natural sound sequence as output using a pre-recorded sound database. A classifier is trained before it is used to recognize the speech from input data. The focus of this paper is recognizing speech from the input data using a classifier.

Our approach is to directly map articulatory position time-series data to phonemes using a classifier without extracting articulatory features such as maximum mouth opening and without calculating the correlation between input data and target acoustics. The input data are time-normalized to fixed width and are sampled to vectors of attributes. An attribute is a time point of articulatory position. Acoustic data obtained in

978-1-4244-4474-8/09/$25.00 ©2009 IEEE

data collection are used for segmenting the input data only. They are not used for recognition. Such an approach would make real-time prediction possible in our articulatory-to-acoustics synthesizer, because there is no computational processing required to extract articulatory features or to calculate the correlation between input data and target sounds in the prediction stage. The classifier is trained before the synthesizer starts running. So it does not matter if the training process is time intensive for our application. The focus of the current research effort was to determine and improve the recognition accuracy.

This investigation has two objectives: (1) To test the accuracy of our recognition approach of directly mapping articulatory position time-series data to phonemes; and (2) To find the best suited classifier for the future development of an articulatory-to-acoustics synthesizer. Three commonly used classifiers, Neural Network (NN), Support Vector Machine (SVM) [1, 2, 16], and Decision Tree (C4.5) [19, 20] were evaluated for their accuracy and processing speed.

II. Experimental Setup

A. Dataset

For this preliminary study, only one single speaker dataset was considered.

The single-speaker dataset contained eight major English vowels /ɑ/, /i/, /ɐ/, /æ/, /ʌ/, /ɔ/, /o/, /u/ produced by a native female college student who was a native American English speaker.

Data collection involved recording 3D motions from the tongue and lips during vowel production. Table I gives a list of the articulators and location from which the 3D movement data were collected along the midsagittal plane. UL refers to the middle point of upper lip and LL refers to the middle point of lower lip. T1, T2, T3 and T4 were placed on the midsagittal line of the tongue. T1 is tongue tip and T4 is tongue back. Only the *y* and *z* coordinates were used in this project, because the movement along the *x* axis is minimal during speech production. Here, *x*, *y* and *z* were defined as spatial dimensions width, height and length in a 3D Cartesian coordinate system.

The position data of each phoneme are time-normalized and sampled to fixed width. The width was fixed, because all classifiers require that input must be in fixed number vectors of attributes. Prior to analysis the movement data were centered about their means.

TABLE I
Articulators (Data Attributes)

Articulator	Location
UL	Upper Lip
LL	Lower Lip
T1	Tongue tip
T2	Tongue Body Front
T3	Tongue Body Back
T4	Tongue Back

B. Classifiers

Three widely used classifiers were used and compared in this study. Neural Networks are a powerful non-linear computational modeling tool, used widely to model the complex relationship between inputs and targets. Support Vector Machine is relatively newer than Neural Networks and has attracted the interests of machine learning researchers in recent years [16]. C4.5 is a decision tree based classification tool [19, 20]. In C4.5, each attribute can be used for splitting the data to smaller sets (branching in the decision tree). Three representative implementations for the classifiers were selected. The commercial implementation of a Feed Forward Back Propagation Network (FFBP) provided in Matlab (MathWorks Inc.), LIBSVM [16], an implementation of "one-against-one" multiple-class classification with default kernel Radial Basis Function (RBF), and C4.5 [19 20] were used in our experiment.

C. Analysis

Cross validation was used to determine the classification and prediction accuracy of each algorithm. Using this method, samples used to train the algorithms are different from those that are used to test their accuracy. This technique is the standard method of testing classification accuracy in machine learning experiments.

In this study, the default settings were used for each classifier. Only the number of units (85) in the hidden layer of the neural network was determined in a preliminary experiment. Prediction probability estimation was enabled in SVM [16]. One reason is that the optimal parameters found in this dataset do not generalize to other datasets, due to the high variation in speech movements. And according to our preliminary experiments, there was no significant difference when some parameters of SVM and C4.5 were changed.

III. Data Collection and Processing

A. Participant

Speech movements were obtained from a native English speaking female college student. The participant had a negative history of hearing or speech impairment.

B. Device

The Electromagnetic Articulograph (EMA) AG500 [15] was used to register the 3D movements of the lip, jaw, and tongue during vowel production. The EMA AG500 records movements by establishing a calibrated electromagnetic field that induces current into small electromagnetic sensor coils that are attached to the articulators. The subject with attached sensors is seated with his/her head within the electromagnetic cube. When the subject speaks, the 3D positions of the sensors are recorded and saved to a desktop computer connecting to

TABLE II
SAMPLE FORMAT

Attributes						Target
ULy1, ULy2, ... ULyn	ULz1, ULz2, ... ULzn	…	T1y1, … T1yn	…	T4z1, ... T4zn	Phoneme

n is equivalent to 10.

the cube. The origin of the coordinate system of EMA is the center of the cube. The anatomically based coordinate system that was used defined the *x* axis as the participant's right and left (horizontal), the *y* axis as the participant's top and bottom (vertical), and *z* axis as the participant's front and back. The spatial accuracy of motion tracking using the EMA (AG500) is approximately 0.5 *mm* [21].

Fig. 1 shows the twelve sensors attached on the subject's head, face, and tongue. Three of the sensors were attached on a pair of glasses the subject wore. HC (Head Center) was on the bridge of the glasses, and HL (Head Left) and HR (Head Right) were on the left and right outside edge of each lense respectively. The movements of HC, HL and HR were used to derive lip and tongue movement data that were independent from head motion. These sensors were attached to a pair of rigid glasses to avoid skin motion artifact [23]. Four of the sensors, T1, T2, T3 and T4, were attached on the midsagittal line of the tongue. There was an interval of around 10*mm* between two continuous tongue sensors [18]. Three of the sensors, JL (Jaw Left), JR and JC, were attached on the canines and one of the incisors. The movements of jaw were prepared for future use only.

C. Stimuli & Procedure

The subject spoke the eight vowels (i.e., /ɑ/, /i/, /e/, /æ/, /ʌ/, /ɔ/, /o/, /u/) in a sequential order at a normal speaking rate. There was a small interval between two continuous vowels. She repeated the sequence twenty times. Seventeen valid sequences of all vowels were obtained in the end. This sample size is comparable to that of other studies [3, 4, 6, 7, 9, 10, 13].

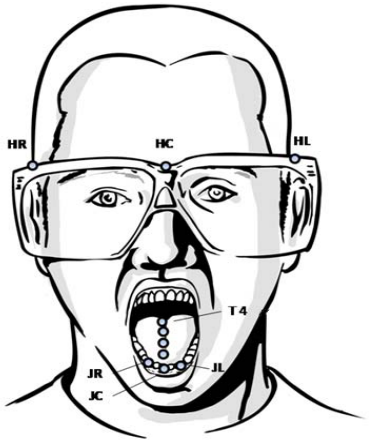

Figure 1. Sensor positions.

D. Data Processing

The movement data were segmented to obtain the movements associated with each vowel, time-normalized, and sampled to vectors which were accepted by the classifiers.

Each movement data file was a recording of a sequence of all eight single vowels, and there were seventeen such recordings. Prior to analysis, the head movements were subtracted from the lip and tongue data. Then a low pass filter of 10 *Hz* was applied to the position data. The acoustic data was used for segmenting the continuous sequence of eight vowels into single vowel segments. The segmentation was done manually.

The articulatory position time-series data for each vowel was downsampled to 10 frames and centered about its mean. All sampled frames of all articulators were concatenated as one sample of 120 (6 articulators × 2 dimensions × 10 frames) attributes as an input sample. According to our preliminary results, 10 time points was sufficient for preserving the details of each movement trace. Table II illustrates how the whole sample was organized. The format was consistent for the three classifiers in our experiment. The three implementations of classifiers had different representations for target/label column in a sample. The Matlab Neural Network used binary encoding, SVM used unique integers, and C4.5 used strings to represent different phonemes.

IV. RESULTS AND DISCUSSION

A. Classification

Six-fold cross validation was conducted for the three classifiers. There were totally six executions. The dataset has totally 136 samples (8 vowels × 17 samples for each vowel). In each execution, three of seventeen samples of each vowel were selected for testing, and the rest were training samples. Since seventeen cannot be divided by six evenly, the first five executions had three samples for each vowel as testing samples, but the last execution has only two samples for each vowel as testing samples. The average recognition rate of the six executions was considered as the cross-validation recognition accuracy.

Table III gives recognition rates for the three classifiers using cross validation. The results are comparable to those of previous research, which have reported recognition accuracies between 70s and low 90s in percentage. In this study, SVM obtained significantly better accuracy than NN and C4.5. C4.5 provided an estimated accuracy in each execution. The average estimated accuracy of C4.5 was 85.48%, which is

TABLE III
RECOGNITION RATE IN CROSS VALIDATION

	NN	SVM	C4.5
Recognition Rate (%)	85.76 (8.78)	91.32 (3.34)	76.07 (11.00)

Standard deviation in parenthesis.

similar to the result of NN.

Fig. 2 provided the classification matrices of the fourth execution in cross validation. The three classifiers had different recognition rates and different erroneous classifications. For example, NN misclassified two /ɐ/s to /ɑ/ and /i/ respectively, and one /ɔ/ to /o/. SVM misclassified one /ɐ/ to /ɑ/. Decision Tree (C4.5) misclassified one /ɐ/ to /ʌ/, and one /ʌ/ to /ɔ/.

	ɑ	i	e	æ	ʌ	ɔ	o	u
ɑ	3							
i		3						
e	1	1	1					
æ				3				
ʌ					3			
ɔ						3		
o	1						2	
u								3

a. Neural Network (87.50%)

	ɑ	i	e	æ	ʌ	ɔ	o	u
ɑ	3							
i		3						
e	1		2					
æ				3				
ʌ					3			
ɔ						3		
o							3	
u								3

b. Support Vector Machine (95.83%)

	ɑ	i	e	æ	ʌ	ɔ	o	u
ɑ	3							
i		3						
e			2		1			
æ				3				
ʌ					2	1		
ɔ						3		
o							3	
u								3

c. C4.5 (91.67%)

Figure 2. Classification matrices in the 4th execution of cross validation.

It should be noted that the results were obtained using the original time-series data without significant or complex data processing. Thus, the recognition accuracy can still be improved. But as we discussed before, the goal of this study is to show the effectiveness and potential of this approach. Improvements and full optimizations is part of our ongoing work.

B. Execution Time

Table IV gives the execution time in the cross-validation in Section IV-B. The training epochs of NN were 7, 9, 7, 7, 7, 7 (average 7.33) respectively in the six executions of cross validation.

SVM and C4.5 were executed on a laptop with 2.1G duo core processor and 2G memory. NN ran on a 2.5G duo core processor with 6G memory desktop, because 2G memory was not enough for NN. NN may need more time in both training and testing stage if running on the same machine as SVM and C4.5, but there may be no significant difference.

Because each algorithm had different implementations (LIBSVM and C4.5 were implemented using C language, and NN was implemented in Matlab), the processing speeds cannot be compared directly. However, the processing time results clearly demonstrate the potential for all three classifiers to be used for near-real time applications. Moreover, because the training will be performed offline before the synthesizer is actually used, the high training cost in time for a NN is not a significant problem.

TABLE IV
EXECUTION TIME (SECOND)

Time	NN	SVM	C4.5
Training	666.72	0.013	0.071
Testing	0.0104	0.005	0.005

C. Decision Trees

C4.5 provides not only the recognition result, but also information on how the recognition (distinguishing vowels) works in the form of a decision tree, which identified the attributes (articulator's coordinate at a time point) used to distinguish vowels.

An example of decision tree generated by C4.5 is given in Fig. 3. The decision tree was used to distinguish the four vowels /ɑ/, /i/, /æ/ and /o/. The decision tree took all 68 (4 vowels × 17 samples for each vowel) samples as input. A path from root to a leaf represents the procedure used to distinguish a given vowel. First, the decision tree distinguished /o/ from the other vowels just by the sixth *z* coordinate of LL (Lower Lip, LLz6). Then, it distinguished /æ/ from the rest by the eighth *z* coordinate of T2 (Tongue Body Front, T2z8). Finally, /ɑ/ and /i/ were distinguished by the fourth *z* coordinate of T3 (Tongue Body Back, T3z4). The numbers in the decision tree in Fig. 3 were normalized articulatory positions in *mm*. The

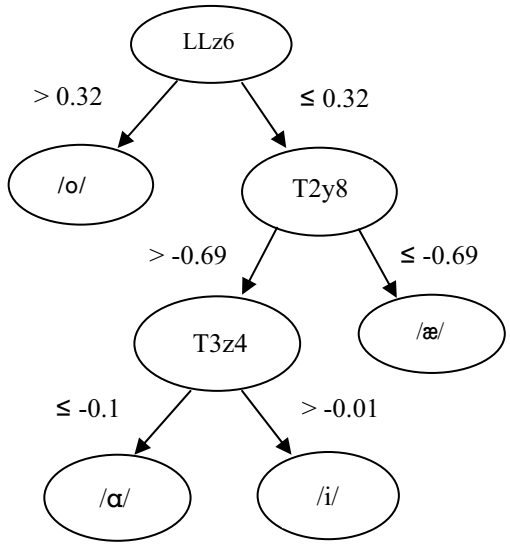

Figure 3. Decision tree for distinguishing /ɑ/, /i/, /æ/ and /ɔ/

tree can distinguish all the four vowels 100% for the given dataset. The decision tree is consistent with the classic descriptions of the distinguishing features among vowels. Furthermore, the decision tree also shows that LL, T2 and T3 contribute more than the rest of the articulators (UL, T1 and T4) in distinguishing the four phonemes /ɑ/, /i/, /æ/ and /ɔ/.

Thus, even though the accuracy of decision tree based classifier was not as good as the others, only this technique provided details about how the time-varying positions were used to distinguish vowels. This information is quite useful and merits further investigation.

V. Summary and Conclusion

A new approach of recognizing speech from articulatory movement was proposed and tested in this paper. The approach directly mapped articulatory position time-series data to vowels using a classifier without extracting articulatory features such as mouth opening. Recognition accuracy and processing speed of three widely used classifiers, Neural Network, Support Vector Machine and Decision Tree (C4.5) were compared working on a single speaker dataset of eight major English vowels /ɑ/, /i/, /ɐ/, /æ/, /ʌ/, /ɔ/, /o/, /u/. Each vowel was spoken seventeen times by a female native English speaker.

The experimental results demonstrated the effectiveness of this approach and its potential for building a real-time articulatory-to-acoustics synthesizer. Cross validation results were 85.76%, 91.32% and 76.07% for the three classifiers respectively. The results were comparable to others presented in the literature. As noted before, the results were obtained with only minimal data preprocessing.

The results showed that our approach is promising. However, additional studies are needed to examine its effectiveness for larger datasets. We briefly list the major conclusions of our study.

(1) Recognition of a limited set of vowels is feasible by mapping articulatory position time-series data to vowels directly using a classifier.

(2) SVM may be the best candidate for building a real-time articulatory-to-acoustics synthesizer among the three classifiers, considering overall performance. SVM obtained significantly better accuracy than NN and Decision Tree in cross validation. In addition, SVM and Decision Tree executed fast either in training or in testing on the given dataset. NN was much slower in training stage but was as comparably fast as SVM and Decision Tree in testing. Thus NN should not be excluded for consideration as a candidate for building the synthesizer.

(3) Decision Tree has its unique advantages. Decision trees give the information how the time-varying articulatory positions are used to distinguish vowels. Decision Tree may be a good candidate for articulatory analysis of speech disorder problems.

VI. Future Work

This research can be extended in the following directions.

(1) Extend dataset to a larger set of vowels, syllables, and words. The extended dataset can include consonant-contexted vowels (CVC), and vowel-contexted consonants (VCV), as well as single vowels. The most interest lies in the word-level articulatory speech recognition.

(2) Continuous speech recognition from articulatory position time-series data. The focus is to recognize not only the phonemes from continuous streams of articulatory movement data but also the intervals (pause) between any two continuous phonemes.

(3) Build a prototype of real-time articulatory-to-acoustics synthesizer. The synthesizer would take articulatory position time-series data as input, recognizes the phoneme/words, and synthesizes natural sound sequences as output using a pre-recorded sound database. Phonemes/words to acoustics synthesis will focus on producing natural sounds.

Acknowledgment

This work was in part funded by the Barkley Trust, Barkley Memorial Center, University of Nebraska-Lincoln and a grant awarded by the National Institutes of Health (R01 DC009890/DC/NIDCD NIH HHS/United States). We would like to thank Dr. Mili Kuruvilla for technical support in data collection.

References

[1] B. E. Boser, I. Guyon, and V. Vapnik. "A training algorithm for optimal margin classifiers," *In Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pp. 144-152. ACM Press, 1992.

[2] C. Cortes and V. Vapnik, "Support-vector network," *Machine Learning*, vol. 20, pp. 273-297, 1995.
[3] K. Shirai, Y. Fukazawa, T. Matzui and H. Matzuura, "A trial of Japanese text input system using speech recognition," *Proc. the 8th Conference on Computational Linguistics Magnetism,* Tokyo, Japan, Sep. 30 – Oct. 04, 1980.
[4] V. S. Sadeghi, and K. Yaghmaie, "Vowel Recognition using Neural Networks," *IJCSNS International Journal of Computer Science and Network Security*, vol. 6, no. 12, pp. 154-158, 2006.
[5] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-Articulator Markov Models for speech recognition," *Automatic Speech Recognition, ISCA Tutorial and Research Workshop (ITRW)*, Paris, France, September 18-20, 2000.
[6] T. Shinchi, Y. Maeda, K. Sugahara, and R. Konishi, "Vowel recognition according to lip shapes by using neural network, " *Proc. of IEEE,* 1998.
[7] P. Cosi, and E. Caldognetto, "Lips and Jaw Movements for Vowels and Consonants: Spatio-Temporal Characteristics and Bimodal Recognition Applications," *NATO ASI Series F Computer And Systems Sciences*, vol. 150, pp. 291-314, 1996.
[8] C. S. Blackburn, and S. J. Young. "Pseudo-articulatory speech synthesis for recognition using automatic feature extraction from X-ray data," *Proc. of the International Conference on Speech and Language Processing*, vol 2, pp 969-972, 1996.
[9] J. Barker and F. Berthommier, "Evidence of correlation between acoustic and visual features of speech," *Proc. International Congress of Phonetic Sciences (ICPhS) '99*, San Francisco, USA, 1999.
[10] J. Jiang, A. Alwan, P. A. Keating, E. T. Auer, and L. E. Bernstein, "On the relationship between face movements, tongue movements and speech acoustics," *Journal on Applied Signal Processing(EURASIP),* vol. 11, pp. 1174–1188, 2002.
[11] C. T. Kello, and D. C. Plaut, "A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters," *Journal of Acoustic Society of America,* vol. 116, pp. 2354-64, 2004.
[12] A. V. Barbosa, and H. C. Yehia, "Measuring the relation between speech acoustics and 2D facial motion," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.
[13] H. C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Linking facial animation, head motion and speech acoustics," *Journal of Phonetics,* vol. 30, pp. 555-568, 2002.
[14] A. Dowd, J. R. Smith, and J. Wolfe, "Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real time," *Language and Speech*, vol. 41 no.1 pp. 1-20, Jan/Mar 1998.
[15] EMA, Carstens Medizinelektronik, GmbH, Germany, http://www.articulograph.de
[16] C. C. Chang and C. J. Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
[17] J. R. Green, L. Phan, I. Nip, and A. Mefferd, "A real-time articulatory controlled vowel synthesizer for research on speech motor learning", *Stem-, Spraak- en Taalpathologie*, 14(Supp.), 46, 2006.
[18] J. Westbury, X-ray microbeam speech production database user's handbook. University of Wisconsin, 1994.
[19] J. R. Quinlan, "C4.5: Programs for Machine Learning," *Morgan Kaufmann Publishers*, 1993.
[20] J. R. Quinlan, "Improved use of continuous attributes in c4.5," *Journal of Artificial Intelligence Research*, 4:77-90, 1996.
[21] Y. Yunusova, J. R. Green, and A. Mefferd, "Accuracy assessment for AG500 electromagnetic articulograph," *Journal of Speech, Language, and Hearing Research,* vol. 52(2) pp. 547-555, 2009.
[22] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 121(2), pp. 723-742, 2007.
[23] J. R. Green, E. M. Wilson, Y. Wang, and C. A. Moore, Estimating mandibular motion based on chin surface targets during speech, *Journal of Speech, Language, and Hearing Research*, vol. 50, pp. 928-939, 2007.