ON THE PHONETIC INFORMATION IN ULTRASONIC MICROPHONE SIGNALS

Karen Livescu

Toyota Technological Institute at Chicago
Chicago, IL 60637, USA
klivescu@uchicago.edu

Bo Zhu, James Glass

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{boz,glass}@mit.edu

ABSTRACT

We study the phonetic information in the signal from an ultrasonic "microphone", a device that emits an ultrasonic wave toward a speaker and receives the reflected, Doppler-shifted signal. This can be used in addition to audio to improve automatic speech recognition. This work is an effort to better understand the ultrasonic signal, and potentially to determine a set of natural sub-word units. We present classification and clustering experiments on CVC and VCV sequences in speaker-dependent and multi-speaker settings. Using a set of ultrasonic spectral features and diagonal Gaussian models, it is possible to distinguish all consonants and most vowels. When clustering the confusion data, the consonant clusters mostly correspond to places and manners of articulation; the vowel data roughly clusters into high, low, and rounded vowels.

Index Terms— Speech recognition, ultrasonic, multimodal

1. INTRODUCTION

A great deal of work has been devoted to the use of non-acoustic signals, especially video [1], in addition to audio for improved speech recognition. Here we consider the signal from an ultrasonic "microphone", a device that emits an ultrasonic sound wave toward the speaker and receives the reflected, Doppler-shifted signal. This is much cheaper than video, both in actual cost and in data rate, and is less intrusive. In this work we investigate the linguistic information in the ultrasonic signal. This is of scientific interest, but also a necessary step for extending the use of ultrasonic signals to larger-vocabulary tasks where sub-word units are used. In this sense, this work can be viewed as an initial attempt at defining ultrasonic sub-word units, analogously to phonemes for audio and visemes for video. Similarly to prior work on video [2, 3, 4], we perform classification and clustering experiments on nonsense consonant-vowel-consonant (CVC) and vowel-consonant-vowel (VCV) sequences, and study their dependence on speaker and phonetic context.

Ultrasonic microphones take advantage of the Doppler effect: When a sinusoidal sound wave at frequency f_0 impinges on a surface moving at velocity v, the frequency of the reflected sound is $f = f_0(1 + \frac{v}{c})$, where c is the speed of sound. The emitted ultrasonic signal in our setup consists of a beam that may impinge on multiple surfaces moving at different velocities, so the reflected signal in general has a complex spectrum. Figure 1 shows example spectrograms of ultrasonic signals from our data collection.

Jennings and Ruck [5] showed early promising results with an ultrasonic dynamic time warping-based "lip-reader" for isolated digits. Zhu *et al.* [6] found that continuous digit recognition in noise benefits significantly from the ultrasonic signal as well. Kalgaonkar *et al.* have studied the use of ultrasonic signals for voice activity detection [7] and speaker recognition [8]. However, to our knowledge no detailed studies have been done on the linguistic information in the signal, analogous to phonetic confusion studies on video.

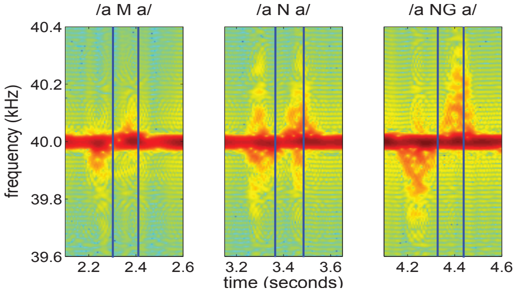


Fig. 1. *Ultrasonic spectrograms for /a M a/, /a N a/, /a NG a/. The vertical lines correspond to boundaries of the nasal consonant.*

Our goal, therefore, is to understand the phonetic information in the ultrasonic signal. Since the signal is generated by articulatory motions, it is plausible that we can discriminate among features such as place of articulation. We may expect, as with video, that we cannot distinguish between voiced and voiceless phonemes. Beyond these general expectations, it is less clear what to expect. It is not clear to what extent we should be able to discriminate among similar articulations in more forward or more back places, e.g. alveolar vs. velar, since the ultrasonic signal is in principle affected only by the velocity of reflecting surfaces. It is also not clear to what extent the signal is affected by speaker and phonetic context.

We follow the rough outline of previous experiments on lip-reading from video [2]. We train classifiers for nonsense utterances containing a set of target vowels and consonants, and analyze their confusions. In the following sections, we describe the ultrasonic hardware, data collection effort, and classification experiments.

2. HARDWARE AND DATA COLLECTION

The ultrasonic hardware is a next-generation version of the one used in [6] and is described in detail in [9]. [1] The ultrasonic transmitter emits a 40 kHz square wave. The reflected signal is received by the ultrasonic receiver, and is then amplified and passed through a 40 kHz bandpass filter and digitized. The audio is simultaneously captured by the on-board or external microphone and low-pass filtered with a cutoff of approximately 8 kHz. Both channels are transferred to a host computer over USB. The device is approximately 1.5 in. high x 2.5 in. wide x 1 in. deep.

This research was supported in part by NIH Training Grant T32 DC000038, and by a Clare Boothe Luce Post-Doctoral Fellowship.

[1] We gratefully acknowledge the assistance of Carrick Detweiler and Iuliu Vasilescu at MIT with the new ultrasonic hardware.

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

Data collection was done in a quiet office environment. Eight speakers, six male and two female, read a script consisting of isolated words each containing a target vowel or consonant. Each speaker was positioned with his/her mouth approximately centered with the ultrasonic sensors and about 6-10" away from the hardware. The audio and ultrasonic channels were simultaneously recorded; here we report on experiments with the ultrasonic signal only. The script consisted of 15 English vowels in the same consonantal environment (/h V d/) and 24 English consonants in four VCV contexts (/a C a/, /i C i/, /u C u/, /ah C ah/), for a total of 111 distinct nonsense words. Two speakers (the first two authors, referred to as Speaker 1 and Speaker 2) recorded twenty sessions of the 111-word script, while the remaining speakers recorded two sessions each.

3. PHONETIC CLASSIFICATION EXPERIMENTS

3.1. Preprocessing and feature extraction

In addition to the reflected ultrasonic signal, the carrier signal is also received directly from the transmitter, and can be strong enough to overwhelm the reflected signal near the carrier frequency. To remove the carrier signal, we first approximate its spectrum as the spectrum of the first frame of each utterance, in which there should be no speech or significant motion. For each remaining frame, we compute a normalized spectrum in which the magnitude at the carrier frequency is matched to the first frame. We subtract the normalized spectrum from the received spectrum, and use the result as the signal for further processing. In addition, each word is segmented semi-automatically: The SUMMIT speech recognizer [10] is used in forced alignment mode to generate two boundaries, one before and one after the target phoneme, and the result is edited manually.

We extract three types of spectral features, the first two of which were used in [6]: (1) **Frequency-band energy averages:** We partition the ultrasonic spectrum into 10 non-linearly spaced sub-bands centered around the carrier frequency, and compute the average energy in each sub-band, in dB relative to the energy at the carrier frequency; (2) **energy-band frequency averages:** We partition the spectrum into 12 *energy* bands and compute the mean frequency in each band; (3) **peak locations:** The ultrasonic spectrograms contain peaks corresponding to forward or backward motions. We use as features the peak times, in particular the times of the maximum and minimum of the frequency average features in a given energy band within a 40ms window of phonetic boundaries.

Next we generate a single vector at each phonetic boundary. We define twelve windows spanning both sides of each boundary (0-6ms, 6-18ms, 18-30ms, 30-60ms, 60-90ms, 90-180ms) and compute the means of the energy and frequency features over each window. Finally, we concatenate the averaged energy and frequency features and the four peak location features (two per boundary) to give a per-utterance feature vector of 532 dimensions. This vector is projected to a smaller dimension using principal components analysis (PCA). The dimensionality for vowel classification tasks was set to maximize the accuracy for each speaker condition; for the consonant tasks, it was set to maximize the mean accuracy over the four contexts for each speaker condition. The PCA dimension ranged from 26 to 52, but did not make a large difference over a wide range.

3.2. Phonetic classification

We use a single diagonal Gaussian to model the distribution of feature vectors for each phonetic class, where a class is a /h V d/ or /V C V/ word. For each test utterance with feature vector O, we classify it by finding the most probable model $C^* = \arg\max_C p(C|O) = \arg\max_C p(O|C)$, where C ranges over vowels or consonants and all classes are equally likely. For the single-speaker experiments, we use 10-fold cross-validation. For each experiment, the data is split into ten non-overlapping subsets, and ten train/test runs are done using a different 90%/10% split in each. We report the average statistics over the ten train/test runs. For the multi-speaker experiments, the same procedure is used with a 13-way split.

Task	Speaker 1	Speaker 2	Multi-speaker
/h V d/	40.7	50.3	33.2
/a C a/	59.4	69.2	47.9
/i C i/	34.8	47.9	30.0
/u C u/	29.8	54.8	29.8
/ah C ah/	55.0	58.3	41.0
all VCV	44.7	57.2	37.2

Table 1. *Accuracies (in %), of vowel (15-way) and consonant (24-way) classification. "All VCV" is the mean accuracy over the four VCV tasks.*

Table 1 shows the overall accuracies. For each cell in the table, a separate set of models was trained on the corresponding data. The multi-speaker condition included all of the recorded data, while the Speaker 1 and 2 conditions included only the corresponding speaker's data. All accuracies are much higher than chance ($\sim 7\%$ for vowels and $\sim 4\%$ for consonants), so there is significant information in the ultrasonic features for these tasks. Second, for the consonant tasks, performance is generally best for the /a/ context, followed by /ah/, /i/, and then /u/. This is expected, since /a/ has the widest lip opening, and therefore the best opportunity for the ultrasonic beam to impinge on surfaces inside the mouth, while the other vowels have progressively narrower lip opening. Third, the highest accuracies are obtained for Speaker 2 and the lowest for the multi-speaker condition. The ultrasonic signal, like the acoustic signal, is therefore quite speaker-dependent.

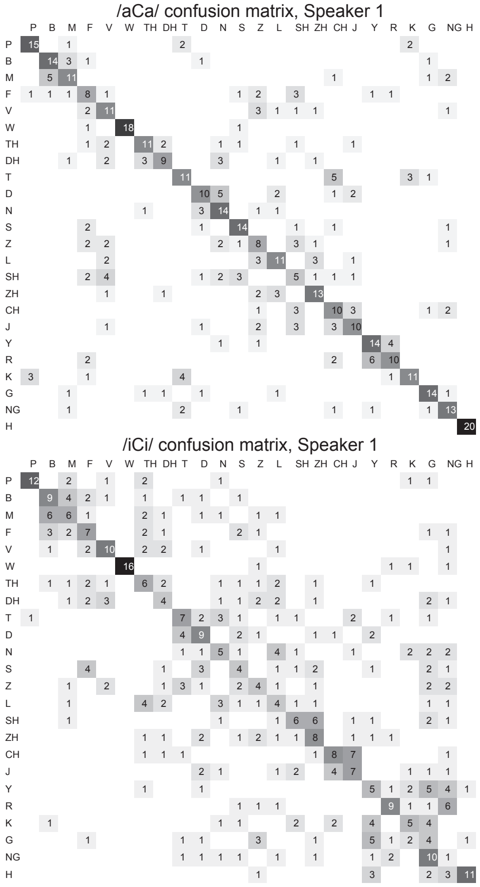
3.3. Confusion matrices

For a better understanding of the misclassifications, we study confusion matrices and clusterings for each task. Here we include a representative subset. Figure 2 shows VCV confusion matrices for Speaker 1 in the /a/ and /i/ contexts. Each matrix cell c_{ij} represents the number of times phone i was classified as phone j, and is displayed numerically and via cell shading. In going from the /a/ to /i/ context, there are more misclassifications, but they tend to cluster around the diagonal, indicating that consonants with similar place are confused (the labels are ordered roughly by place). For the /u/ context (not shown), the off-diagonal confusions are much more uniformly distributed, while the case of /ah/ is similar to that of /a/.

Figure 3 shows the overall VCV confusion matrices (summed over the four contexts) for Speaker 2 and the multi-speaker case, and the vowel confusion matrix for Speaker 1 (excluding "extreme" diphthongs /ay/, /oy/, /aw/). Many of the consonant confusions are expected, such as between voiced/voiceless pairs. However, this is not always the case, e.g. /p/ and /b/ are rarely confused. This may be because of the difference in voice onset time, and therefore peak locations. For Speaker 2, the confusions are concentrated near the diagonal, indicating within-place confusions. This also holds for Speaker 1 (not shown), and less strongly in the multi-speaker case. In Speaker 1's vowel data, the confusions are more concentrated among vowels with similar front/back position.

3.4. Clustering: The search for ultrasonic sub-word units

Finally, we attempt to better understand how the phonemes cluster using hierarchical clustering. The main questions here are (1) is there a need to cluster phonemes into ultrasonic sub-word units, analogously to visemes for lipreading, i.e. are there some phonemes that cannot be distinguished from the ultrasonic signal, and (2) if we



/aCa/ confusion matrix, Speaker 1

	P	B	M	F	V	W	TH	DH	T	D	N	S	Z	L	SH	ZH	CH	J	Y	R	K	G	NG	H
P	15		1						2												2			
B		14	3	1						1												1		
M		5	11														1					1	2	
F	1	1	1	8	1							1	2		3				1	1				
V				2	11								3	1	1	1							1	
W				1		18						1												
TH				1	2		11	2			1	1			1			1						
DH			1		2		3	9			3			1		1								
T									11								5				3	1		
D										10	5			2			1	2						
N							1			3	14		1	1										
S				2						1		14			1		1						1	
Z				2	2						2	1	8		3	1							1	
L					2								3	11		3		1						
SH				2	4					1	2	3			5	1	1	1						
ZH					1			1					2	3		13								
CH													1		3		10	3				1	2	
J					1					1			2		3		3	10						
Y											1		1						14	4				
R				2													2		6	10				
K	3			1					4											1	11			
G			1				1	1		1				1								14	1	
NG			1						2			1					1		1			1	13	
H																								20

/iCi/ confusion matrix, Speaker 1

	P	B	M	F	V	W	TH	DH	T	D	N	S	Z	L	SH	ZH	CH	J	Y	R	K	G	NG	H
P	12		2		1		2				1										1	1		
B		9	4	2	1		1		1	1		1												
M		6	6	1			2	1		1	1		1	1										
F		3	2	7			2	1				2	1									1	1	
V		1		2	10		2	2		1				1									1	
W						16							1							1	1		1	
TH		1	1	2	1		6	2			1	1	1	2		1			1					
DH			1	2	3			4			1	1	2	2		1						2	1	
T	1								7	2	3	1		1	1			2		1		1		
D									4	9		2	1			1	1		2					
N									1	1	5	1		4	1			1			2	2	2	
S				4				1		3		4		1	1	2			1			2	1	
Z			1		2			1	3	1		2	4	1		1						2	2	
L			1				4	2			3	1	1	4	1	1						1	1	
SH			1								1			1	6	6		1	1			2	1	
ZH							1	1		2		1	2	1	1	8		1	1	1				
CH							1	1	1							1	8	7					1	
J										2	1			1	2		4	7			1	1	1	
Y							1			1									5	1	2	5	4	1
R												1	1	1						9	1	1	6	
K		1									1	1			2		2		4		5	4		
G				1					1	1			3			1			5	1	2	4		1
NG									1	1	1	1		1		1			1	2		10	1	
H													1						3			2	3	11

Fig. 2. *VCV confusion matrices in two contexts for Speaker 1.*

wish to cluster the phonemes, what are the natural clusterings induced by the data? For this purpose, the confusion matrices provide us with a natural notion of inter-phone dissimilarity. We represent each phone i by its row vector of confusion frequencies, $c_{ij} \forall j$, and use a measure of dissimilarity between distributions as the dissimilarity between phones. As in previous work on video [3], we use the ϕ measure, a symmetric and normalized relative of χ^2: $\phi = \sqrt{\left(\chi^2(i,j) + \chi^2(j,i)\right)/2N}$, where N is the number of tokens of each phone and $\chi^2(i,j)$ is the χ^2 statistic comparing the confusion frequencies of phoneme i to those of phoneme j. We then cluster hierarchically using average linkage: At each iteration, the two clusters with the smallest mean ϕ between their members are merged. The results of clustering the overall VCV confusions for Speaker 2 and the multi-speaker set, and Speaker 1's vowel confusions, are shown in Figure 4. The y-axis corresponds to ϕ.

First, we address the question of whether it is necessary to cluster the phonemes at all. Are there any phoneme pairs that cannot be distinguished at better than chance level? If each phoneme i is characterized by its confusion frequencies, $c_{ij} \forall j$, then we can use a χ^2 goodness of fit test to test the null hypothesis that phoneme i's confusions are drawn from the same distribution as phoneme j's. By this measure, at a significance level of 0.05, all consonant pairs in the multi-speaker data are distinguishable, and the only indistinguishable vowel pair is {aa, ao}.

Second, we look for natural clusterings: What clusterings would give highly discriminable classes? It is arguable what "highly discriminable" means; here we look at the result of clustering the consonants into 10 classes and the vowels into 5 classes. In Figure 4, the corresponding level is marked in each dendrogram with a dotted red line. Table 2 shows the resulting clusters. The consonant clusters often correspond to places of articulation (e.g., {b m}, {g ng}), but sometimes to manners (e.g., {t k}, {y r}). The vowel clusters correspond roughly to high, low, and rounded.

Task	Clusters
Sp. 1 VCV	{b m} {p} {w} {f th s sh} {v dh z zh} {d n ch j} {l} {t k} {y r g ng} {h}
Sp. 2 VCV	{b m} {p} {w} {f v dh l g ng} {th s z sh zh ch j} {d n} {t} {k} {y r} {h}
Multi-sp. VCV	{b m} {p} {w} {f v th dh s z sh zh} {d n l} {ch j} {t k} {g ng} {y r} {h}
Sp. 1 vowels	{iy ih ey eh ah} {ae aa ao} {er} {uh} {uw ow}
Sp. 2 vowels	{iy ih eh ah} {ey ae aa ao} {er ow} {uh} {uw}
Multi-sp. vowels	{iy ey} {ih eh ah} {ae aa ao} {er uw ow} {uh}

Table 2. *Clusters derived from the dendrograms in Figure 4.*

4. CONCLUSIONS

We have studied phonetic discrimination in ultrasonic microphone signals, in the setting of nonsense /h V d/ and /V C V/ classification, and can draw some initial conclusions. First, we can discriminate at above chance level among all consonant pairs in the combined VCV data, and among most vowel classes. It may therefore not be necessary to group consonants into equivalence classes, as is done for video. The experimental setup is idealized, however; this analysis should be extended to continuous speech. The phonetic confusions differ between speakers and phonetic contexts. For example, consonants in an /i/ or /u/ context are more difficult to distinguish than those in an /a/ or /ah/ context. It remains to be seen whether different features or more complex models (e.g. Gaussian mixtures), trained on more data, would be more robust to such variation.

From confusion matrices and hierarchical clustering, we have found that the most salient groupings of consonants include both place and manner of articulation classes, and do not necessarily include voiced/voiceless pairs. This differs from video, where the most salient divisions are along place of articulation. This may be because the ultrasonic microphone is sensitive mainly to the velocity, and not position, of reflecting surfaces. When clustering the multi-speaker consonant data into ten classes, the resulting clusters are {{b m} {p} {w} {f v th dh s z sh zh} {d n l} {ch j} {t k} {g ng} {y r} {h}}. When clustering the multi-speaker vowel data into five classes, the result is {{iy ey} {ih eh ah} {ae aa ao} {er uw ow} {uh}}.

This study is an initial step toward understanding the phonetic information in ultrasonic signals, and toward the question of what a good set of ultrasonic sub-word units (if any) may be. This will help us to expand the use of ultrasonic signals beyond the limited domains in which they have been used so far. Future work includes investigation of additional ultrasonic features and direct comparisons of phonetic discrimination using ultrasonic and video signals.

5. REFERENCES

[1] G. Potamianos, C. Neti, G. Gravier, and A. Garg, "Automatic recognition of audio-visual speech: Recent progress and challenges," *Proc. IEEE*, vol. 91, no. 9, 2003.

[2] J. Xue, J. Jiang, A. Alwan, and L. E. Bernstein, "Consonant confusion structure based on machine classification of visual features in continuous speech," in *Audio-Visual Speech Processing Workshop*, 2005.

[3] J. Jiang, E. T. Auer Jr., A. Alwan, P. A. Keating, and L. E. Bernstein, "Similarity structure in visual speech perception and optical phonetic signals," *Perception and Psychophysics*, vol. 69, no. 7, pp. 1070–1083, 2007.

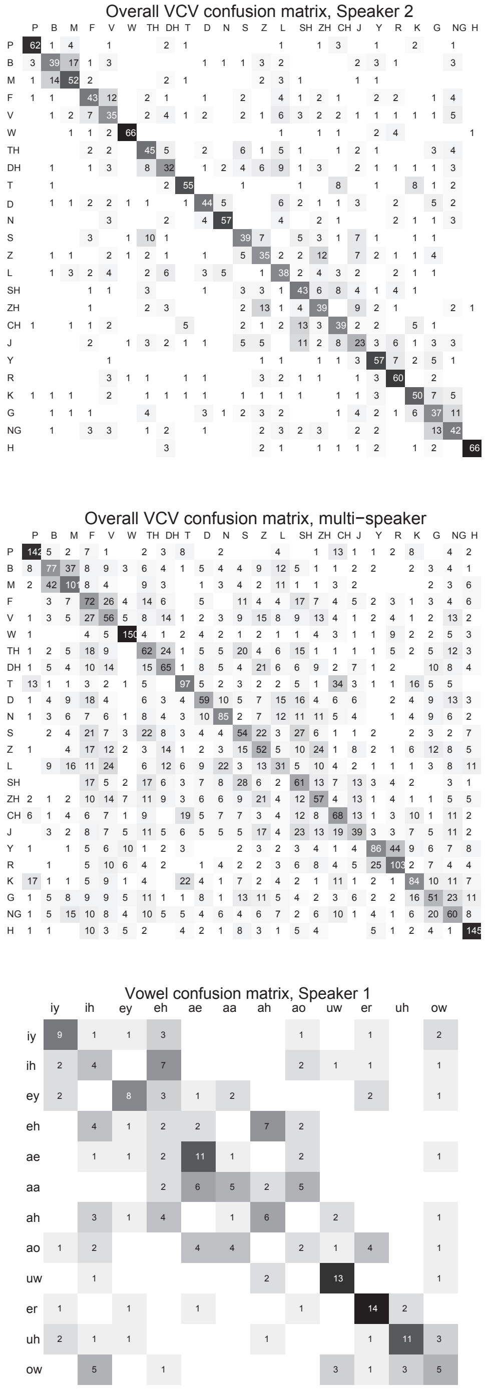


Fig. 3. *Overall confusion matrices for VCV and vowel tasks.*

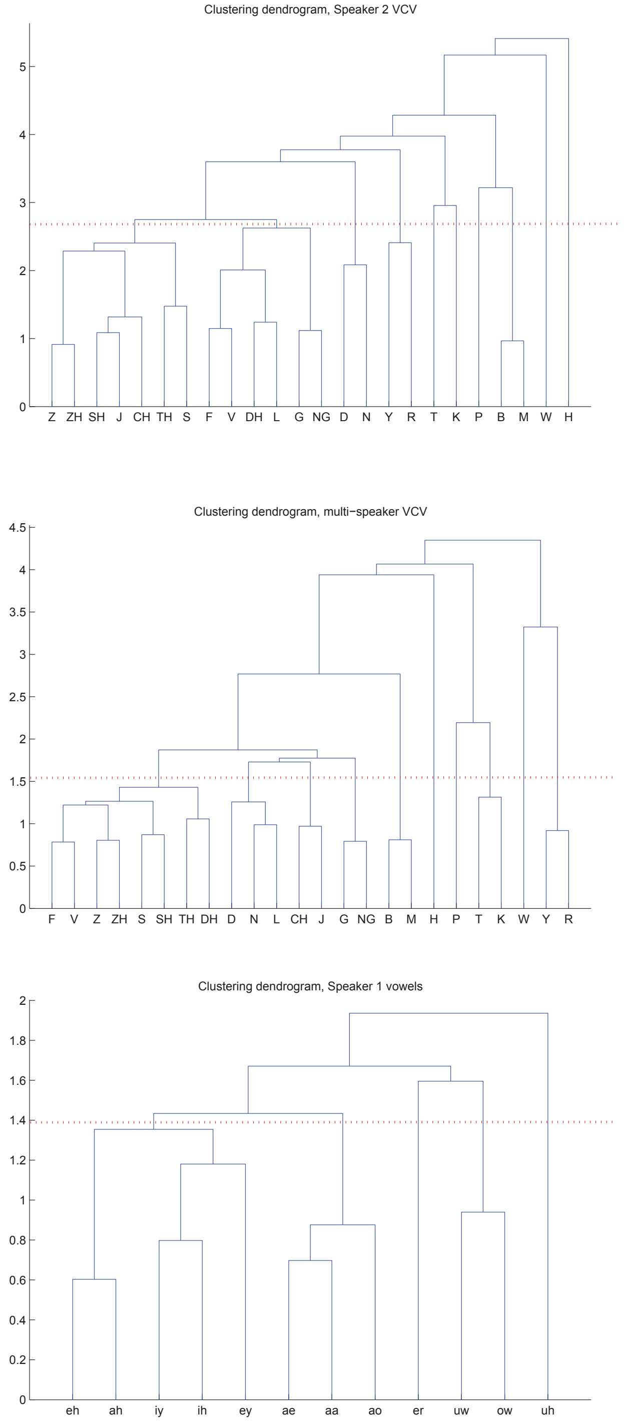


Fig. 4. *Clustering dendrograms corresponding to Figure 3.*

[4] B. E. Walden, R. A. Prosek, A. A. Montgomery, C. K. Scherr, and C. J. Jones, "Effects of training on the visual recognition of consonants," *J. Sp. Hearing Res.*, vol. 20, pp. 130–145, 1977.

[5] D. L. Jennings and D. W. Ruck, "Enhancing automatic speech recognition with an ultrasonic lip motion detector," in *ICASSP*, 1995.

[6] B. Zhu, T. J. Hazen, and J. R. Glass, "Mutimodal speech recognition with ultrasonic sensors," in *Interspeech*, 2007.

[7] K. Kalgaonkar, H. Rongquiang, and B. Raj, "Ultrasonic Doppler sensor for voice activity detection," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 754–757, 2007.

[8] K. Kalgaonkar and B. Raj, "Ultrasonic Doppler sensor for speaker recognition," in *ICASSP*, 2008.

[9] C. Detweiler and I. Vasilescu, "Ultrasonic speech capture board: Hardware platform and software interface," Indep. study final paper, 2008.

[10] J. Glass, "A probabilistic framework for segment-based speech recognition," *Comp. Sp. Lang.*, vol. 17, pp. 137–152, 2003.