# Comparing Human and Machine Recognition Performance on a VCV Corpus

*Odette Scharenborg*[1] *and Martin Cooke*[2]

[1]Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
[2]Speech and Hearing Research Group, Dept. of Computer Science, University of Sheffield, UK
O.Scharenborg@let.ru.nl, M.Cooke@dcs.shef.ac.uk

## Abstract

Listeners outperform ASR systems in every speech recognition task. However, what is not clear is where this human advantage originates. This paper investigates the role of acoustic feature representations. We test four (MFCCs, PLPs, Mel Filterbanks, Rate Maps) acoustic representations, with and without 'pitch' information, using the same back-end. The results are compared with listener results at the level of articulatory feature classification. While no acoustic feature representation reached the levels of human performance, both MFCCs and Rate maps achieved good scores, with Rate maps nearing human performance on the classification of voicing. Comparing the results on the most difficult articulatory features to classify showed similarities between the humans and the SVMs: e.g., 'dental' was by far the least well identified by both groups. Overall, adding pitch information seemed to hamper classification performance.

**Index Terms**: human-machine comparison, acoustic feature representations, articulatory feature classification.

## 1. Introduction

Listeners outperform automatic speech recognition (ASR) systems at every level of speech recognition, including the very basic level of consonant recognition (e.g., [1],[2]). However, humans are generally able to exploit knowledge about the world, the environment, the topic of discourse, and use other information that is unavailable to an ASR system. Another difference is the way humans and computers represent the input they receive: humans are able to use all information that is present in the acoustic signal, while ASR systems can only use the information that is encoded in the acoustic features. Ideally, acoustic features preserve all information relevant for the automatic recognition of speech, but in reality this is not always the case.

There have been relatively few studies comparing human and automatic speech recognition on the same task (for an overview see [3]). However, there are many insights which might be gained by carrying out detailed comparisons. Furthermore, knowledge about what information (or which cues) is present in the speech signal and is most robustly detected by listeners (usually tested in adverse listening conditions; e.g., [4],[5]) can be used to improve machine recognition performance.

In this paper, we systematically test various acoustic feature representations using the same back-end and compare the results with listeners' results on the same task in order to explore the question of which information or cues are necessary (in the feature representation) for human and automatic speech recognition. We used support vector machines (SVMs) as our back-end systems since they can generalise to small amounts of high-dimensional data. We analyse the recognition results in terms of (mis)detection of articulatory features (AFs).

## 2. Experimental set-up

### 2.1. Support Vector Machines

SVMs are binary maximum margin classifiers (for an introduction, see [6]). The principle underlying SVMs is the *maximum margin principle*. Given two separable classes, the decision boundary is found by maximising the margin or distance between the two classes such that no data occupy the space in-between. The decision boundary is chosen so that it is geometrically furthest away from both classes. The vectors near the decision boundary are the support vectors (SVs).

When the data is non-separable, a soft margin is used that allows some points to enter the margin or be misclassified entirely. Incursions into the margin are penalised, so a search for the best solution maximises the margin and minimises the penalties simultaneously.

In our experiments, we used the LIBSVM package [7], which achieves multi-class classification by error correcting codes. The RBF kernel was used for the experiments reported in this paper.

### 2.2. Acoustic feature representations

The acoustic feature representations that were compared are all based, to some extent, on the human auditory system:

- Mel Frequency Cepstral Coefficients (MFCC).
- Perceptual Linear Predictive coefficients (PLP, [8]).
- Mel Filterbanks (Fbank).
- Rate maps (RMaps, [9]).

The first three acoustic feature representations were created using HTK [10]. In order to create the Fbank feature vectors, the speech signal was parameterised using the energy in 24 Mel Filters augmented with c0, and the 1st and 2nd derivatives. The MFCC, PLP, and Fbank feature vectors were based on 25 ms windows and a 10 ms frame shift. For the PLP and MFCC feature vectors, the speech was parameterised with coefficients c0…c12 and also augmented with 1st and 2nd derivatives. RMaps were created by filtering the signal into 40 bands using gammatone filters with centre frequencies equally-spaced on an ERB-rate scale from 50 to 8000 Hz. The instantaneous envelope at the output of each filter was smoothed (time constant 8 ms) and downsampled to 10 ms frames to produce an auditory spectral representation. Note that for the creation of the RMaps no 1st and 2nd derivatives were used.

Following [11], all features were extended with a context window of ±3 frames in order to take into consideration the dynamic nature of speech, which usually spans more than the duration of one frame. All acoustic feature representations were also tested with added pitch information. Pitch was extracted from the speech signal using Praat [12]. For every frame, a single pitch value $F$ in Hz was computed, which was subsequently converted to a semitone value $ST$ using the following formula:

$$ST = 12 \bullet \log 2\left(\frac{F}{50}\right)$$

The semitone value was then added to the acoustic feature vector. Unvoiced frames were assigned a random $F$ value between 5 and 30 ($F$=0 was also tested but resulted in worse results than the method proposed here). The acoustic representations without pitch have the suffix ‘$_-$’, while the suffix ‘$_+$’ denotes that pitch information has been used.

## 2.3. Material

The speech material was recorded in an IAC single-walled acoustically isolated booth at the University of Sheffield. Twenty eight native English talkers (age 18-49; 12F, 16M) produced vowel-consonant-vowel (VCV) tokens in isolation by reading out tokens presented on a computer screen. Each talker produced each of the 24 consonants (/b, d, g, p, t, k, s, sh, f, v, th, dh, t, z, zh, h, dj, m, n, ng, w, r, y, l/) in nine vowel contexts consisting of all possible combinations of the three vowels /i:/ (as in "beat"), /u:/ (as in "boot"), and /ae/ (as in "bat"). Each VCV was produced using both initial and final stress (e.g. 'aba vs ab'a).

The resulting ‘VCV corpus’ was split into a training (8F, 8M), a development (4M), and an independent test set (4F, 4M). After removing unusable tokens identified during post-processing, the training set consists of 6,664 clean tokens. The development set consists of 192 clean tokens. The test set contains 16 instances of each of the 24 consonants, for a total of 384 tokens. For more information on the corpus, see [13].

## 2.4. Segmentation and selection of VCV data

In order to train and evaluate the SVM classifiers, the VCV data needs to be labelled at the frame level. To that end, 30 HMM models were trained using HTK: 24 consonants and two models for each of the three vowels (one to model the initial and one to model the final vowel context of the VCV). The speech was parameterised using the earlier described MFCCs. Each of the models consisted of 3 emitting states with 32 Gaussian mixtures, while the silence model used 64 mixtures. The models were subsequently retrained on a set of 91 hand-segmented VCVs taken from the VCV training set.

For evaluation, this model set was then used to segment the 91 sound files of which a manual segmentation was available. The average difference in number of frames in the position of the boundary between the manual segmentation and the automatically derived segmentation was 2.4 for the silence-V boundary, 1.5 for the V-C boundary, and 2.8 for the C-V boundary. Finally, a segmentation for all data sets was obtained using forced alignment.

The SVMs were only trained and tested on the frames corresponding to consonants. Removing the vowel and silence frames from the three sets resulted in a training set consisting of 73,488 frames, a test set of 3,918, and a development set of 2,321 frames.

## 2.5. Articulatory features

We use three AFs and their respective ‘values’ to compare the acoustic representations amongst themselves and with human performance. The SVM training, development, and test data sets were created by replacing the frame-level phonemic labels with the canonical AF value as listed in Table 1.

## 2.6. Human data

Twenty four native English listeners who reported no hearing problems identified the 384 VCVs of the test set. Perception tests ran under computer control in the IAC booth. Listeners were presented with a screen on which the 24 consonants were represented using both ASCII symbols and with an example word containing the sound. Listeners were phonetically-naive and were given instructions as to the meaning of each symbol. They underwent a short practice session prior to the main test.

*Table 1. Mapping of phonemes on the AF values.*

| **Consonant** | **manner** | **place** | **voice** |
|---|---|---|---|
| *b* | plosive | labial | +voice |
| *d* | plosive | alveolar | +voice |
| *g* | plosive | velar | +voice |
| *p* | plosive | labial | -voice |
| *t* | plosive | alveolar | -voice |
| *k* | plosive | velar | -voice |
| *s* | fricative | alveolar | -voice |
| *sh* | fricative | (pre)palatal | -voice |
| *f* | fricative | labial | -voice |
| *v* | fricative | labial | +voice |
| *th* | fricative | dental | -voice |
| *dh* | fricative | dental | +voice |
| *ch* | affricate | (pre)palatal | -voice |
| *z* | fricative | alveolar | +voice |
| *zh* | fricative | (pre)palatal | +voice |
| *h* | fricative | glottal | -voice |
| *dj* | affricate | (pre)palatal | +voice |
| *m* | nasal | labial | +voice |
| *n* | nasal | alveolar | +voice |
| *ng* | nasal | velar | +voice |
| *w* | glide | labial | +voice |
| *r* | liquid | (pre)palatal | +voice |
| *y* | glide | (pre)palatal | +voice |
| *l* | liquid | alveolar | +voice |

*Table 2. The AF accuracy and the percentage of the training data that are SVs for the acoustic representations.*

| **Ac. Feat.** | **manner** | | **place** | | **voice** | |
|---|---|---|---|---|---|---|
| | %SV | %Acc | %SV | %Acc | %SV | %Acc |
| $\text{MFCC}_-$ | 39.4 | 92.2 | 57.6 | 84.4 | 24.7 | 96.1 |
| $\text{MFCC}_+$ | 33.5 | 91.7 | 57.0 | 82.1 | 26.1 | 95.8 |
| $\text{PLP}_-$ | 39.9 | **93.2** | 58.3 | 82.9 | 25.9 | 95.8 |
| $\text{PLP}_+$ | 33.9 | 93.0 | 64.0 | 84.4 | 26.8 | 95.6 |
| $\text{Fbank}_-$ | 55.9 | 67.0 | 65.7 | 58.7 | 34.9 | 86.8 |
| $\text{Fbank}_+$ | 69.6 | 70.9 | 74.2 | 55.1 | 28.4 | 89.4 |
| $\text{RMaps}_-$ | 21.7 | 90.1 | 35.9 | **85.5** | 11.9 | **96.6** |
| $\text{RMaps}_+$ | 23.3 | 90.1 | 36.1 | 85.2 | 15.4 | 95.3 |
| Human | N/A | 98 | N/A | 97 | N/A | 97 |

*Table 3. Values of the $\gamma$ and c parameters for each SVM.*

| **Ac. Feat.** | **param** | **manner** | **place** | **voice** |
|---|---|---|---|---|
| $\text{MFCC}_-$ | *c* | 5 | 10 | 10 |
| | $\gamma$ | 0.01 | 0.01 | 0.5 |
| $\text{MFCC}_+$ | *c* | 5 | 10 | 10 |
| | $\gamma$ | 0.005 | 0.01 | 0.5 |
| $\text{PLP}_-$ | c | 5 | 10 | 5 |
| | $\gamma$ | 0.005 | 0.01 | 0.5 |
| $\text{PLP}_+$ | *c* | 5 | 5 | 5 |
| | $\gamma$ | 0.005 | 0.01 | 0.5 |
| $\text{Fbank}_-$ | *c* | 100 | 100 | 100 |
| | $\gamma$ | 0.1 | 0.05 | 1 |
| $\text{Fbank}_+$ | *c* | 200 | 100 | 100 |
| | $\gamma$ | 0.005 | 0.01 | 0.5 |
| $\text{RMaps}_-$ | *c* | 50 | 10 | 100 |
| | $\gamma$ | 0.005 | 0.01 | 0.5 |
| $\text{RMaps}_+$ | *c* | 50 | 50 | 50 |
| | $\gamma$ | 0.005 | 0.01 | 0.1 |

# 3. Classification results

For each acoustic representation a set of SVM classifiers was trained: one SVM for each AF. The classifiers then performed an AF classification where each frame was assigned an AF value. Table 2 shows the classification results for each AF separately for each of the acoustic representations and the human listeners. Bold figures indicate the best scoring acoustic representation for that AF. The SVM classification accuracy is calculated as follows: the AF value most often occurring over all frames belonging to one consonant is taken as the recognition result. The accuracy is then the total number of correct AF values divided by 384 (consonants). Scores for listeners were calculated as follows: summary confusion matrices (i.e. summed over all listeners) were processed to produce scores for *manner*, *place*, and *voice* as well as for the individual AF values. For example, any 'plosive' recognised as a 'plosive' was treated as correct for the 'plosive' feature value. Similarly, any consonant whose *manner* was correctly recognised was treated as correct for the *manner* score. Overall, listeners achieved 93% correct consonant identification but their scores for *manner*, *place* and *voice* were near to ceiling, i.e. close to 100% correct.

Table 2 also lists the number of support vectors as a percentage of the amount of training data. The percentage of SVs indicates the SVM complexity: more SVs suggest either more complex decision boundaries or more overlapping data. For completeness, Table 3 lists the values of the $\gamma$ and $c$ parameters in the SVMs. These values were estimated on the development set. The $\gamma$ is the reciprocal of the RBF kernel width squared: a large $\gamma$ implies narrower RBFs. $c$ sets the amount of regularisation, i.e., simpler decision boundaries vs. fitting the training data. If $c$ is large then the SVM constructs more complex decision boundaries to better fit the training data but may result in poor generalisation.

Table 2 shows that the AF *voice* is recognised best for all acoustic representations, followed by *manner* and then *place*. The percentage SVs shows a reversed trend: *voice* has the lowest percentage SVs, while *place* has the highest. This is not surprising: classification of *voice* is a binary task while classifying *manner* and *place* involves making a choice from six AF values. The values for $\gamma$ for *place* are small for all acoustic representations, which indicates that the width of the RBFs is reasonably large. This suggests that the clusters representing the *place* AF values are not highly localised, but that there is considerable overlap between the *place* AF value clusters, resulting in poorer generalisation than for the other six-valued AF *manner*.

In accordance with earlier results in the literature (e.g., [1],[2]), humans outperform machines for all three AFs, although for *voice* the difference is not significant (all mentions of significance are at the 95% level and calculated using a *t*-test), except for Fbank (-/+ pitch). Comparing the classification scores of the eight acoustic representations shows that the RMaps$_-$ twice has the highest score, however overall, MFCC$_-$ features have the best score, with the RMaps$_-$ coefficients as a close second. PLP$_-$ perform only slightly worse than RMaps$_-$ coefficients. Fbank$_-$, however, perform (significantly) the worst of all tested acoustic representations. The addition of pitch information led to a small drop in performance for MFCC, PLP, and RMaps features (apart from *place* for PLP). However, it was beneficial for Fbank for both *manner* and *voice*. The differences were not significant.

Table 4 shows the accuracies for each AF value separately for all acoustic representations and the listeners. The '-'/'+' columns indicate whether pitch information has been added to the acoustic representations. Again, the bold figures indicate the best scoring acoustic representation for that AF value.

PLP (-/+ pitch) had the highest accuracies for most *manner* values, while RMaps$_+$ had the highest AF value accuracies for *place*. For *manner*, the order of the AF value classification results for all acoustic representations are fairly similar: 'fricative' and 'plosive' are the top 2 best recognised (except for Fbank$_-$: 'fricative' and 'nasal' are the top 2 best recognised). It is difficult to compare these results with the human results, though, because humans perform near ceiling for 'plosive', 'fricative', 'glide', and 'liquid'. The 'liquid' feature is worst classified for all acoustic representations apart from Fbank$_+$ ('glide') and RMaps$_-$ ('nasal'). This is not entirely in accordance with human performance where 'affricate' is the most difficult AF value to recognise.

The two best classified *place* AF values are 'palatal' and 'labial'. Humans scored near perfect for 'alveolar', 'velar', and 'glottal' with slightly lower scores for 'palatal' and 'labial'. Like humans, all acoustic classifiers (except Fbank (-/+ pitch)) score (significantly) lowest for 'dental'. Despite 'dental' being the hardest *manner* value to classify ([11]), RMaps classified 'dental' significantly better than the other acoustic representations.

'-voice' is systematically classified better than '+voice' by Fbank (-/+ pitch; both significant) and RMaps (-/+ pitch, not significant), as for humans. The difference in classification accuracy for '-voice' and '+voice' for MFCC$_+$ is also significant.

# 4. Discussion

Four acoustic feature representations, with and without additional 'pitch' information, were compared with listeners at the level of identification of traditional AFs. While no representation reached the levels of human performance, MFCCs, RMaps$_-$ and PLPs achieved good scores. For example, the RMaps$_-$ representation correctly identified the *voice* feature 96.6% of the time, compared to 97% for listeners. Listeners significantly outscored SVMs for both *manner* and *place*, the latter by 11.5 percentage points. For specific AF values, there are some similarities in performance between the two groups: for instance, 'dental' *place* was by far the least well identified.

*Table 4. AF value classification accuracies per acoustic representation and 'for the listeners; -/+ indicates whether pitch information has been used.*

| **AF value** | **Accuracy (%)** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *MFCC* | | *PLP* | | *Fbank* | | *RMaps* | | *Human* |
| | - | + | - | + | - | + | - | + | |
| *manner* | | | | | | | | | |
| plosive | 96.9 | 94.8 | **97.9** | **97.9** | 72.9 | 76.0 | 93.8 | 92.7 | 99 |
| fricative | 96.5 | 96.5 | **97.2** | 96.5 | 75.0 | 81.3 | **97.2** | 95.8 | 98 |
| affricate | 87.5 | 87.5 | **96.9** | 90.6 | 62.5 | 65.6 | 84.4 | 87.5 | 92 |
| nasal | 85.4 | **87.5** | 85.4 | 85.4 | 66.7 | 79.2 | 77.1 | 81.3 | 97 |
| glide | **87.5** | 84.4 | 84.4 | **87.5** | 56.3 | 34.4 | 84.4 | 84.4 | 99 |
| liquid | 81.3 | 81.3 | 81.3 | **84.4** | 31.3 | 40.6 | 81.3 | 81.3 | 99 |
| *place* | | | | | | | | | |
| labial | **94.8** | 89.6 | 88.5 | 91.7 | 79.2 | 76.0 | 88.5 | 88.5 | 95 |
| dental | 53.1 | 53.1 | 56.3 | 56.3 | 28.1 | 15.6 | **65.6** | **65.6** | 86 |
| alveolar | 83.3 | 82.3 | 81.3 | **85.4** | 57.3 | 57.3 | 83.3 | 82.3 | 99 |
| palatal | 90.6 | 89.6 | 92.7 | 91.7 | 63.5 | 63.5 | **96.9** | 92.7 | 97 |
| velar | 81.3 | 81.3 | 81.3 | 83.3 | 47.9 | 35.4 | 81.3 | **87.5** | 99 |
| glottal | 68.8 | 56.3 | 62.5 | 56.3 | 12.5 | 6.3 | 68.8 | **75.0** | 98 |
| *voice* | | | | | | | | | |
| +voice | 96.7 | **97.1** | 96.7 | 96.7 | 85.0 | 87.5 | 96.7 | 95.0 | 97 |
| -voice | 95.8 | 94.4 | 95.1 | 94.4 | 90.3 | 93.1 | **97.2** | 96.5 | 98 |

It is worth noting that the task for the humans and the SVMs was not identical: listeners were tested on the task of intervocalic *consonant* identification, i.e., *one decision* per consonant; while the SVMs assigned an *AF value* to each frame of the test material, i.e., *multiple decisions* per stretch of speech belonging to one consonant. Despite this difference, we believe that the results are comparable since listeners' results have been processed to become AF value detection scores. However, one difference remains: human listeners had full access to the vowel information and thus information from coarticulation, while the SVMs were trained and tested only on the consonantal information and thus only had vowel information in the ±3 frame windows. In this light, the scores obtained by the SVMs for certain feature representations are surprisingly high.

The local nature of the SVM decision might explain why *place* and *manner* are less well identified than *voice*. It is conceivable that voicing information is available throughout the segment, while *place* can be expected to be strongly influenced by coarticulation. In fact, the 'dental' value was mis-identified as 'labial' on 44% of occasions by both MFCC-based approaches compared to 11% for listeners. This confusion is attributable to the similarities between the 'dental' consonants /th, dh/ and the 'labial' consonants /f, v/, respectively. The similarity in production of the sounds (all involve the teeth) results in similar relatively flat spectra for both groups. The main difference between the two groups is the position of F2, which is higher for 'dental' [14]. Likewise *manner* information is not necessarily uniquely discriminable at all points, especially for consonants such as 'plosive', 'affricate' and 'glide'. Therefore, it will be interesting to investigate the AF classification performances for all acoustic representations using classifiers that do not use frame-based classification.

Perhaps surprisingly, adding additional information about pitch hampers classification performance slightly for MFCCs, RMaps, and PLPs in most conditions. It is clear that other cues to voicing are available since these features perform very well at the voicing distinction without the additional information, and it is possible that the other voicing cues are more reliable than the explicit use of a pitch estimate. Furthermore, theoretically, pitch information could help in distinguishing two ambiguous classes if one class consisted solely (or mainly) of voiced and the other of unvoiced samples, because of the different distributions of voiced and unvoiced frames. However, most AF values contain both '+voice' and '-voice' samples, e.g,, '+voice' and '-voice' dentals both occur in the 'dental' set. So when carrying out frame-based AF classification, pitch is not needed to discriminate between classes. Still, pitch might be beneficial if it could be applied to longer segments, for instance during consonant recognition.

Of the four acoustic feature representations, FBank stands out for its poor performance. It is intriguing that the other filterbank representation, RMaps, performed significantly better. There are two crucial differences between the two. First, RMaps have better frequency resolution in the region below 1 kHz. Second, the temporal resolution of RMaps varies with frequency, and is particularly fine in the higher frequencies. The contribution of these factors to better performance on this task requires further study.

## 5. Conclusions

Four acoustic feature representations, with and without additional 'pitch' information, were compared with listeners on a task of articulatory feature classification. While no representation reached the levels of performance of human listeners, both MFCCs and Rate maps achieved good scores, with Rate maps nearing human performance on the classification of voicing. The comparison of the machine and human results on the articulatory features that were most difficult to classify showed that there is again some agreement: e.g., 'dental' was by far the least well identified by both humans and machines.

The similarities in performance for some of the articulatory feature (values) between the two groups suggest that the information encoded in the acoustic representations (except for Fbank coefficients) and the information available to human listeners is similar. However, we will further investigate this issue and the question which cues in the speech signal are most robustly detected by listeners by comparing machine and human performance in adverse conditions in follow-up research.

## 6. Acknowledgements

This research was supported by a Veni-grant from the Netherlands Organisation for Scientific Research (NWO) to the first author. Both authors were supported by Marie Curie Project "Sound to Sense". The first author would like to thank Bert Cranen (Radboud University Nijmegen) for discussions on the acoustic representations.

## 7. References

[1] Lippmann, R., "Speech recognition by machines and humans", Speech Comm., 22 (1), 1997, 1-15.

[2] Meyer, B., Wesker, T., Brand, T., Mertins, A., Kollmeier, B., "A human-machine comparison in speech recognition based on a logatome corpus", Proc. Speech Recog. and Intrinsic Variation, Toulouse, France, 2006.

[3] Scharenborg, O., "Reaching over the gap: A review of efforts to link human and automatic speech recognition research", Speech Comm., 49, 2007, 336-347.

[4] Cooke, M., "A glimpsing model of speech recognition in noise", JASA, 119 (3), 2006, 1562-1573.

[5] Sroka, J., Braida, L., "Human and machine consonant recognition", Speech Comm., 45, 2005, 401-423.

[6] Burges, C.J.C., "A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery, 2 (2), 1998, 1-47.

[7] Chang, C.-C., Lin, C.-J. "LIBSVM: a library for support vector machines", Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[8] Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech", JASA, 84 (4), 1990, 1738-1752.

[9] Cooke, M.P., Modelling auditory processing and organization. Cambridge, UK: Cambridge University Press, 1993.

[10] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., The HTK book (for HTK version 3.2). Techn. Report, Cambridge University, Eng. Dept., 2002.

[11] Scharenborg, O., Wan, V., Moore, R.K., "Towards capturing fine phonetic variation in speech using articulatory features", Speech Comm., 49, 2007, 811-826.

[12] Boersma, P. Weenink, D., Praat: doing phonetics by computer (Version 4.6.02), http://www.praat.org/, 2007.

[13] http://www.odettes.dds.nl/challenge_IS08/

[14] Kent, R.D., Read, Ch., The Acoustic Analysis of Speech. London-San Diego: Whurr Publishers - Singular Publishing Group, 1992.