

FREQUENCY SELECTIVE FILTERING OF THE MODULATION SPECTRUM AND ITS IMPACT ON CONSONANT IDENTIFICATION

Thomas U. Christiansen and Steven Greenberg
Centre for Applied Hearing Research
Technical University of Denmark

Thomas Ulrich Christiansen
DTU Centre for Applied Hearing Research, Ørsted-DTU
Ørstedes Plads Bygning 352,
2800 Lyngby
Denmark
tuc@oersted.dtu.dk

Abstract

The spectro-temporal coding of Danish consonants was investigated using an information-theoretic approach. Listeners were asked to identify eleven different consonants spoken in a CV[*l*] syllable context (where C refers to the initial consonant, V refers to one of three vowels, [i, a, u], and [*l*] refers to the syllable-final liquid segment). Each syllable was processed so that only a portion of the original audio spectrum was present. Narrow (three-quarter octave) bands of speech, with center frequencies of 750 Hz, 1500 Hz and 3000 Hz, were presented individually and in combination with each other. The modulation spectrum of each band was low-pass filtered at 24, 12, 6 and 3 Hz. Confusion matrices of the consonant-identification data were computed, and from these the amount of information transmitted for each of three phonetic feature dimensions – voicing, manner and place of articulation – was calculated for each condition. This form of analysis provides a simple means of determining whether information associated with each phonetic feature dimension combines linearly across the audio spectrum, and, if not, delineates a method for characterizing the (non-linear) nature of information integration. In addition, the analysis provides a means to associate specific portions of the modulation spectrum with phonetic feature properties. Such analyses indicate that:

- (1) Accurate, robust decoding of place-of-articulation information requires broadband cross-spectral integration
- (2) Place-of-articulation information is associated most closely with the modulation spectrum above 6 Hz, with the most significant contribution coming from the region above 12 Hz.
- (3) Place-of-articulation information is crucial for accurate consonant recognition. Hence, consonant decoding requires cross-spectral integration of the modulation spectrum above 8 Hz.
- (4) Voicing is mainly associated with the modulation spectrum between 3 and 6 Hz (with a smaller contribution made by the region above 12 Hz).
- (5) Manner of articulation is most closely associated with the portion of the modulation spectrum above 12 Hz.

This form of information-theoretic analysis can be used to delineate those parts of the speech signal of greatest importance for encoding phonetic features associated with intelligibility and speech understanding.

Introduction and Overview

Despite decades of research, the acoustic cues important for consonant identification are not well understood. Traditionally, consonants have been described primarily in spectral terms. The distribution of energy across the acoustic frequency axis has long been considered the primary determinant of a segment's consonantal identity (e.g., [12]).

In contrast, the consonant's temporal properties are often ignored. However, a variety of studies indicate that low-frequency modulations play a crucial role (e.g., [5] [7] [8] [9]). Modulations below 16 Hz are considered particularly important for understanding speech [5]. Frequencies above 16 Hz may also contribute important information under certain conditions [2][7][11]. Currently lacking is a detailed understanding of how modulation information is combined across the acoustic frequency spectrum, as well as a delineation of how spectral and temporal cues interact in decoding the speech signal.

This study investigates the spectro-temporal cues associated with identification of Danish consonants through systematic filtering of the modulation spectrum at different parts of the audio frequency spectrum. Because of speech's inherent redundancy, much of the signal's audio frequency content was discarded in order to compel listeners to focus on restricted regions of the spectrum. The modulation pattern associated with each spectral region was systematically modified (low-pass filtered) and the results of this manipulation evaluated in terms of consonant identification and the amount of information transmitted in terms of each segment's constituent phonetic features.

Decomposing consonants into their constituent phonetic features is important for understanding the auditory mechanisms used to decode speech. The phonetic dimensions of voicing (e.g., differentiating the consonants, [p, t, k] from [b, d, g]¹), articulatory manner (e.g., distinguishing [b] from [m]) and place of articulation (e.g., distinguishing [p] from [t] and [k]) can be used to quantitatively assess the contribution of each spectral and modulation region to consonant identification. This is done by computing confusion matrices and calculating the amount of information transmitted for each phonetic feature dimension under a wide range of spectro-temporal conditions. Because a range of spectral regions is represented in the signal conditions presented, it is possible to ascertain the contribution of each part of the spectrum to consonant identification when presented individually and in combination with other spectral regions. Such information can ultimately be used to enhance the design of future-generation hearing aids.

Experimental Stimuli

Stimuli were Danish monosyllabic words and nonsense syllables originally recorded at Aalborg University [4]. Each test word was preceded by a short carrier phrase "På pladsen mellem hytten og..." The subject's task was to identify the initial consonant of each test token. The stimuli contained one of eleven initial consonants, [p, t, k, b, d, g, m, n, f, s, v], followed by one of three vowels, [i, a, u].

¹ We use the term voicing because it is conventionally used for distinguishing the English stop consonants. The term "aspiration" is the conventional phonetic term for the Danish counterparts. The naming has no bearing on the results.

Each token concluded with the liquid segment [l] (e.g., “talle,” “tulle,” “tille”). The material was spoken by two talkers (one male, one female). The speech signals were recorded with a high-quality microphone in a sound-insulated environment; their original sampling rate was 20 kHz, at which the signal processing was performed. Subsequently the speech signals were up-sampled to 44.1 kHz for stimulus presentation.

Signal Processing

Each test token was processed in the following manner. The acoustic frequency spectrum was partitioned into three separate channels (“slits”), each three-quarters of an octave wide (and were computed using a 256-point FFT and a sampling rate of 44.1 kHz). The lowest slit was centered at 750 Hz, the middle slit at 1500 Hz and the highest slit at 3000 Hz. Each slit was presented either in isolation or in combination with one or two other slits.

Each slit was low-pass modulation filtered using software developed by Les Atlas and colleagues, known as the “Modulation Toolbox” [3]. The low-pass cutoff frequency of modulation filtering ranged between 24 Hz and 3 Hz in the following steps: 24 Hz, 12 Hz, 6 Hz, 3 Hz. These parameters span much of the range of low-frequency modulation information typically encountered in spoken material. Only a single slit was modulation filtered for any given stimulus. In one set of conditions, each slit was presented in its original (i.e., no modulation filtering) form. The carrier phrase was always presented in its original (i.e., unfiltered) form.

Subjects

Six individuals (5 male, 1 female) between the ages of 22 and 28 participated in the study. All reported normal hearing and no history of audiological pathology. Subjects were compensated for their time.

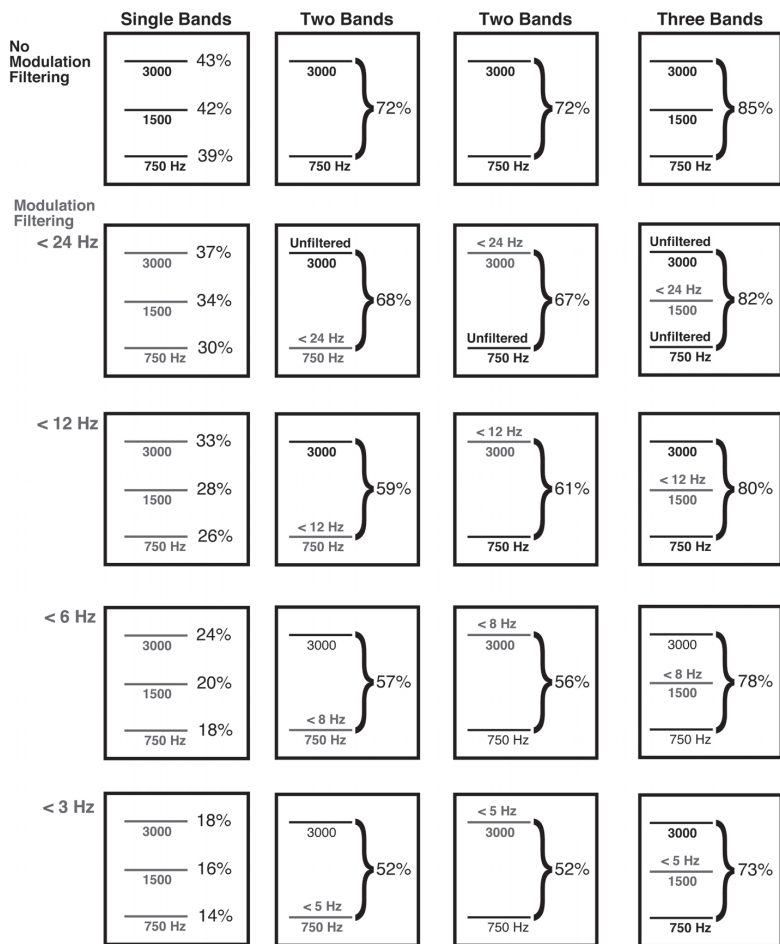
Stimulus Presentation

Experimental stimuli were presented diotically over Sennheiser HD-580 headphones at a sound pressure level of 65 dB SPL to subjects seated in a double-walled sound booth. After each stimulus presentation, the subject selected one of eleven consonants displayed on a computer monitor. No feedback regarding consonant identification accuracy was provided.

The stimulus conditions were:

1. Slits presented individually – 750 Hz, 1500 Hz, 3000 Hz
2. Slits presented in combination with others – 750 + 3000 Hz, 750 + 1500 Hz, 1500 + 3000 Hz, 750 + 1500 + 3000 Hz
3. Low-pass modulation filtering – < 24 Hz, < 12 Hz, < 6 Hz, < 3 Hz, no modulation filtering (as a control)
4. Unfiltered (in either audio frequency or modulation; i. e., the original unprocessed syllables)

Stimuli in each condition were presented three times (for each talker). Altogether, there were 2244 stimulus presentations for each subject. Total testing time per listener was approximately six hours (including brief breaks every 20 minutes or so) administered over a period of two or three separate days.



Conditions marked in gray were low-pass modulation filtered

Figure 1: Consonant identification accuracy (in terms of percent correct) for each experimental condition (average of six subjects). The original (unprocessed) syllables were correctly identified 98.8% of the time.

Data Analysis

The data were analyzed in several different ways. Accuracy of consonant identification was computed for all conditions (Figure 1), both in terms of average and individual performance. Because individual subject variability was relatively small, only average data are reported in this study.

In addition, consonant identification was scored in terms of how well consonantal phonetic properties – voicing, manner and place of articulation – were recognized. When a consonant is correctly identified, its constituent phonetic features are (by definition) also accurately decoded. However, when a consonant is incorrectly identified, it is rare that all of its constituent phonetic features are also inaccurately decoded. Usually, one or two of the features is correctly recognized. This can be deduced from perceptual confusion matrices (see Table I).

Information Transmission Analysis – A Primer

Consonant perception is normally studied in terms of accuracy for individual segments. Because consonants are systematically related to each other (in terms of phonetic features), scoring only in terms of the proportion of consonants correct may obscure patterns associated with auditory mechanisms pertaining to cross-spectral integration and modulation analysis. Confusion matrices of consonantal identification error patterns provide a straightforward means of delineating how much information associated with constituent features is transmitted.

		RESPONSE											
		p	t	k	b	d	g	s	f	v	n	m	
STIMULUS	P	19	13	1	0	0	0	0	3	0	0	0	
	S	1	32	2	0	1	0	0	0	0	0	0	
	T	1	8	27	0	0	0	0	0	0	0	0	
	I	0	0	0	25	10	0	0	0	0	0	1	
	M	0	0	0	2	34	0	0	0	0	0	0	
	U	0	0	1	5	7	22	0	0	1	0	0	
	L	0	0	0	0	0	0	32	4	0	0	0	
	U	0	1	0	0	1	0	23	10	1	0	0	
	S	0	0	0	0	0	1	0	0	27	1	7	
		n	0	0	0	0	0	0	0	0	1	32	3
		m	0	0	0	0	0	0	0	0	8	4	24

Table I: Average confusion matrix for the 750 + 3000 Hz condition (without modulation filtering), where consonant recognition performance is 72% correct.

A representative confusion matrix (associated with the condition “Two bands, No modulation filtering” in Figure 1) is shown in Table I. Rows in this matrix represent the consonants presented (“Stimulus”), while columns represent the subjects’ responses. If a consonant were always identified correctly the score in that segment’s cell would be “36.” In the matrix shown, no consonant is identified perfectly. One consonant, [d], is identified correctly 34 times. However, there are 19 instances when a different consonant is also identified as [d]. In order to compute the “true” amount of information associated with each segment, a bias-neutral metric (such as d' used in signal detection theory) is required. Because we are interested in computing the information associated with constituent phonetic features associated with each consonant, we use information transmission (as defined by Miller and Nicely [10]) as the recognition metric.

In order to compute the amount of information transmitted, the eleven consonants of the recognition set are partitioned into three (overlapping) groups on the basis of the phonetic properties of voicing, articulatory manner and place of articulation, as illustrated in Table II. Voicing refers to the presence (or absence) of glottal vibration. Manner refers to the mode of articulatory production (stop, nasal, fricative) and place of articulation refers to the locus of articulatory constriction (anterior, medial, posterior). Voicing is a binary distinction, while manner and place engender three class distinctions.

	VOICING	MANNER	PLACE
p	0	0	0
t	0	0	1
k	0	0	2
b	1	0	0
d	1	0	1
g	1	0	2
s	0	1	1
f	0	1	0
v	1	1	0
n	1	2	1
m	1	2	0

Table II: Phonetic features for the 11 consonants used in the current study. Voicing is a binary feature dimension, while Manner and Place are ternary feature dimensions.

Confusion matrices for each phonetic-feature dimension can be derived from the original confusion matrix by summing the results for each feature group. In essence, each phonetic-feature dimension is treated as an independent information channel. For example, the consonant confusion matrix illustrated in Table I can be decomposed into three separate feature confusion matrices, as shown in Table III.

		RESPONSE		
		Anterior	Medial	Posterior
S	PLACE			
	Anterior	125	53	2
	Medial	11	131	2
	Posterior	7	15	50
T		Place = 77.3% correct		
I	MANNER	Stop	Fricative	Nasal
M	Stop	211	4	1
U	Fricative	3	97	8
L	Nasal	0	9	63
U		Manner = 93.7% correct		
S	VOICING	Voiced	Unvoiced	
	Voiced	215	1	
	Unvoiced	3	77	
		Voicing = 99.0% correct		

Table III: Phonetic-feature confusion matrices for the same data shown in Table I. Note the disparity in feature recognition performance.

Although a consonant may be identified incorrectly, there may be information pertaining to its constituent properties correctly decoded. Information about voicing is generally decoded accurately even when the consonant is poorly recognized (Figure 2). Information pertaining to manner of articulation is often correctly decoded when the consonant is incorrectly identified (Figure 2). Place of articulation is usually incorrectly decoded when a consonant is not correctly recognized (Figure 2). For a consonant to be accurately identified, voicing, manner and place information must all be correctly decoded. What these analyses demonstrate is that consonant identification depends largely on identifying the place-of-articulation dimension correctly.

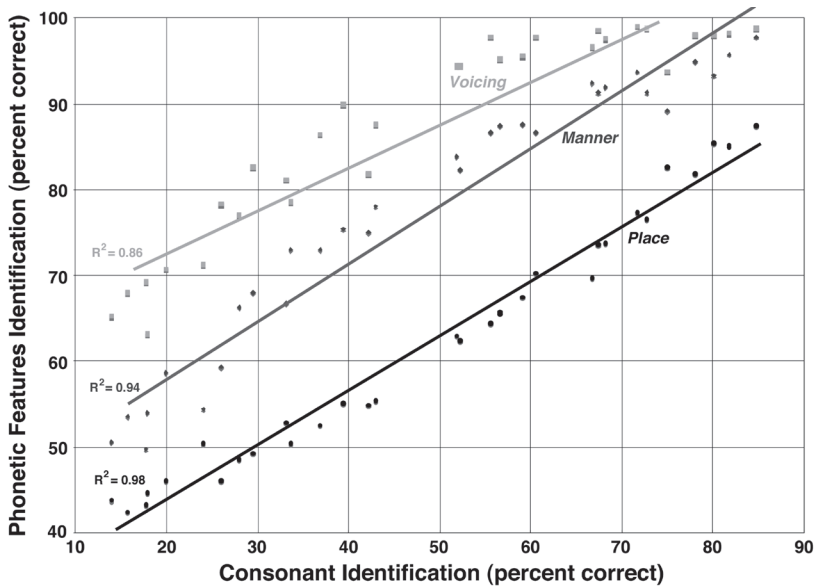


Figure 2: Consonant identification accuracy as a function of phonetic feature classification for the same conditions and listeners depicted in Figure 1. The correlation coefficient (R^2) is shown for each.

In order to compute the amount of information associated with a particular feature and stimulus condition it is necessary to calculate the co-variance between a specific stimulus and response category (this is done to neutralize the effect of response bias). The information associated with Voicing, Manner and Place is computed using equation (1):

$$T(c) = - \sum_{i,j} p_{ij} \log \frac{P_i P_j}{P_{ij}} \quad (1)$$

where $T(c)$ refers to the number of bits per feature transmitted across channel c , p_{ij} is the probability of feature, i , co-occurring with response j , p_i is the probability of feature, i , occurring and p_j is the probability of response j occurring. The amount of information transmitted in the example above is 0.91 bits for voicing, 1.09 bits for manner and 0.60 bits for place.

Information Transmission of Phonetic Features

When the data are plotted in terms of information transmitted, interesting patterns emerge (Figure 3). Information combines differently across the audio spectrum for each phonetic-feature dimension. Manner combines relatively linearly; the information associated with two slits is approximately double that of one, while three slits provide nearly three times the information (for the conditions where no slit is modulation filtered). Voicing combines linearly for two slits, but no further information is gained with the addition of a third slit (i.e., saturation). Place of articulation combines synergistically (two or three slits are much better than one). Figure 4 examines cross-spectral integration in greater detail.

When individual slits are low-pass modulation filtered, there is a progressive decline in the amount of information transmitted for each feature. The pattern is different when two or three slits are presented concurrently. For three slits (where only a single slit is low-pass modulation filtered) there is relatively little degradation of information transmitted. For voicing, there is no degradation whatever, consistent with the observation that voicing information requires only two slits to reach a maximum level. Manner-of-articulation information degrades slightly with low-pass modulation filtering, while place information degrades somewhat more.

Information Transmission of Phonetic Features

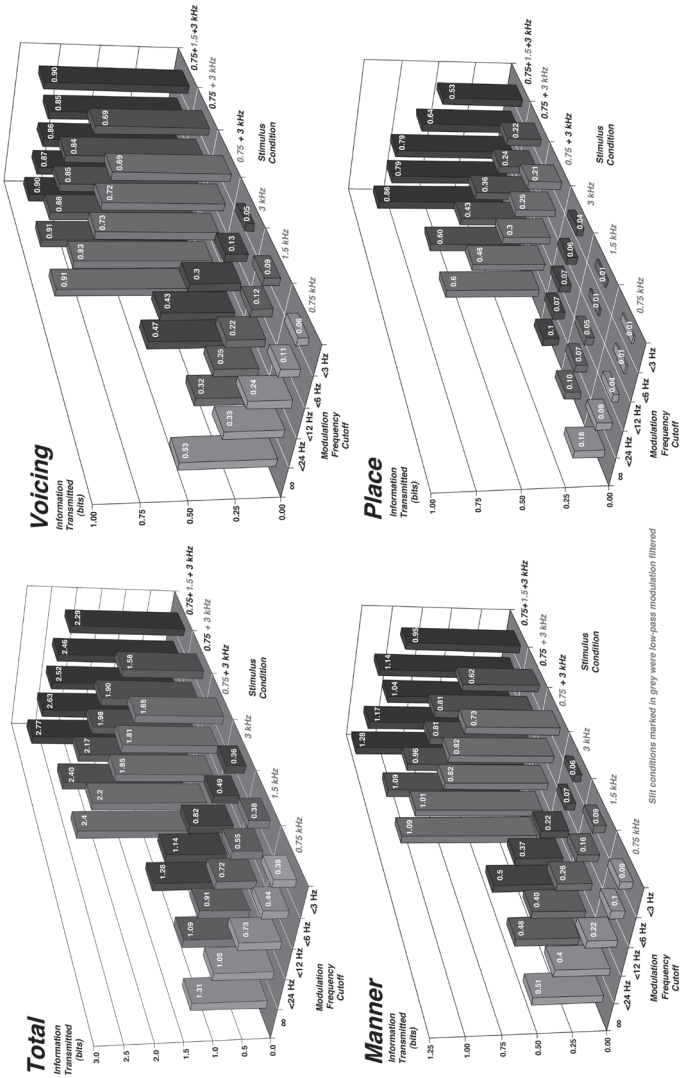


Figure 3: Amount of information transmitted for each phonetic feature (and for all features together – “Total”) as a function of spectral slit and low-pass modulation filtering condition. Data points reflect averages from six listeners.

Place of articulation is the phonetic feature that most depends on cross-spectral integration. Under virtually all stimulus conditions, there is substantially greater than linear summation across slits. Moreover, the amount of information transmitted within any single slit is relatively small (substantially less than either manner or voicing). In other words, place information requires a broad span of speech to be decoded accurately. Place of articulation is also the feature most closely associated with the ability to accurately decode consonantal identity (Figure 2). From such data, we conclude that:

- (1) Accurate, robust decoding of place-of-articulation information requires broadband cross-spectral integration
- (2) Place of articulation information is associated most closely with the modulation spectrum above 6 Hz with the most significant contribution coming from the region above 12 Hz.
- (3) Place of articulation information is crucial for accurate consonant recognition. Hence, consonant decoding requires cross-spectral integration of the modulation spectrum above 6 Hz.
- (4) Voicing is mainly associated with the modulation spectrum between 3 and 6 Hz (with a smaller contribution made by region above 12 Hz).
- (5) Manner of articulation is most closely associated with the modulation spectrum above 12 Hz.

Thus, there is both an acoustic and a perceptual basis for distinguishing among phonetic features associated with the relevant regions of the modulation spectrum and cross-spectral integration of such information.

Conclusions and Significance

Speech is highly redundant. Conventional methods of estimating the contribution made by different parts of the audio spectrum [1] and the modulation spectrum [9] fail to dissociate these two dimensions. Nor do they examine the specific contribution made by cross-spectral integration to speech decoding in a quantitative way.

The current study provides a systematic means by which to measure the contribution of each portion of the audio spectrum and modulation spectrum to human speech processing, both in isolation and in combination with each other (as well as with other sources of information, e.g., visual, speech-reading cues) [6] [7].

Such methods could also be used to model the speech decoding process for application to future-generation hearing aid design as well as other speech-based technologies.

Acknowledgments

Support for this study was provided by Forskningsrådet for Teknologi og Produktion (TUC) and Danmarks Tekniske Universitet (SG). The authors thank Professor Torsten Dau for his encouragement and support, and Professor Les Atlas for advice concerning the use of the “Modulation Toolbox” developed in his laboratory at the University of Washington, Seattle. Steven Greenberg served as a visiting professor at the Technical University of Denmark during the time this project was performed. He is with Silicon Speech, Santa Venetia, CA (USA).

References

- [1] Allen, J.B. (1994) How do humans process and recognize speech? *IEEE Trans. Speech Audio Proc.* 2: 567-577.
- [2] Apoux, F. & Bacon, S.P. (2004) Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise. *J. Acoust. Soc. Am.* 116: 1671-1680.
- [3] Atlas, L., Li, & Q. Thompson, J. (2004) Homomorphic modulation spectra. *Proc. Int. Conf. Audio, Speech & Signal Proc. (ICASSP)*, pp. 761-764.
- [4] Chan, D., Fourcin, A., Gibbon, D., Granstrom, B., Huckvale, M., Kokkinakis, G., Kvale, K., Lamel, L., Lindberg, B., Moreno, A., Mouropoulos, J., Senia, F., Trancoso, I., Veld C., & Zeiliger, J. (1995) EUROM – A Spoken Language Resource for the EU. Proceedings of the 4th European Conference on Speech Communication and Speech Technology (Eurospeech-95). pp. 867-870
- [5] Drullman, R., Festen, J.M. & Plomp, R. (1994) Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.* 95: 2670-2680.
- [6] Greenberg, S. (2005) A multi-tier theoretical framework for understanding spoken language. In *Listening to Speech: An Auditory Perspective* (S. Greenberg and W. A. Ainsworth, eds.). Mahwah, NJ: Erlbaum, pp. 411-433.
- [7] Greenberg, S. & Arai, T. (2004) What are the essential cues for understanding spoken language? *IEICE Trans. Inf. & Syst.* E87-D:1059-1070.
- [8] Greenberg, S., Arai, T. & Silipo, R. (1998) Speech intelligibility derived from exceedingly sparse spectral information, *Proceedings of the 5th International Conference on Spoken Language Processing*, pp. 74-77.
- [9] Houtgast, T. & Steeneken, H.J.M. (1985). A review of the MTF-concept in room acoustics, *J. Acoust. Soc. Am.* 77: 1069-1077.
- [10] Miller, G. A. & Nicely, P. E. (1955) An analysis of perceptual confusions among some English consonants. *J. Acoustic. Soc. Am.* 27: 338-352.
- [11] Silipo, R., Greenberg, S. & Arai, T. (1999) Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations, *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech-99)*, pp. 2687-2690.
- [12] Stevens, K. N. (1998) *Acoustic Phonetics*. Cambridge, MA: MIT Press.

