

DISTINGUISHING SPECTRAL AND TEMPORAL PROPERTIES OF SPEECH USING AN INFORMATION-THEORETIC APPROACH

Thomas U. Christiansen and Steven Greenberg

Centre for Applied Hearing Research, Technical University of Denmark, Kgs. Lyngby, Denmark
tuc@oersted.dtu.dk; steveng@silicon-speech.com

ABSTRACT

The spectro-temporal coding of Danish consonants was investigated using an information-theoretic analysis. Listeners identified eleven consonants spoken in CV[l] context. In each condition, only a portion of the original spectrum was played. Center frequencies of 750, 1500 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 the phonetic features – voicing, manner and place of articulation – was calculated. From these analyses we conclude that:

- (1) Decoding place-of-articulation information requires significant cross-spectral integration
- (2) Place of articulation depends on modulations above 6 Hz, and is *crucial* for consonant recognition
- (3) Voicing requires modulations between 3 and 6 Hz.
- (4) Manner depends on modulations greater than 12 Hz.

Keywords: Speech perception, information theory, modulation spectrum, consonant identification.

1. INTRODUCTION

Which acoustic cues are most important for understanding spoken language? Traditionally, the speech signal has been analyzed and described primarily in spectral terms. In contrast, temporal properties have been largely ignored. However, there is increasing evidence that low-frequency energy fluctuations play a crucial role, particularly those below 16 Hz (e.g., [2][4][5]). Modulations higher than 16 Hz may also contribute under certain conditions [1][3][8]. Currently lacking is a detailed understanding of how low-frequency amplitude-modulation cues are combined across the acoustic frequency spectrum, as well as how spectral and temporal information *interact*. Such

knowledge may enhance our understanding of how spoken language is processed in noisy and reverberant environments by both normal and hearing-impaired individuals.

2. EXPERIMENTAL METHODS

The current study investigates the spectro-temporal cues associated with identification of Danish consonants through filtering of the modulation patterns in different regions of the audio (frequency) spectrum. Because of speech's inherent redundancy, much of the signal's audio frequency content must be discarded in order to delineate the interaction between spectral and temporal information.

The amplitude modulations associated with each of three separate spectral regions were low-pass filtered independently and the resultant signal processing evaluated in terms of consonant identification and the amount of information associated with each consonant's constituent phonetic features. The phonetic features voicing, articulatory manner, and place of articulation were used to assess the contribution of each audio-frequency channel and modulation-frequency region to consonant recognition. This was done by computing confusion matrices and calculating the amount of information transmitted for each phonetic feature. In this way, the contribution of each acoustic frequency region to consonant recognition could be discerned when presented alone and in tandem with other spectral bands.

Stimuli were Danish monosyllabic words and nonsense syllables recorded in a sound-insulated environment at Aalborg University (Denmark). The sampling rate was 20 kHz (up-sampled to 44.1 kHz for presentation). The acoustic frequency spectrum was partitioned into three separate channels ("slits"), each three-quarters of an octave wide. The lowest slit was centered at 750 Hz, the middle one at 1500 Hz and the highest slit at 3000 Hz. Each slit was presented either in isolation or in

Table 1: Consonant identification accuracy (percent correct) for each condition (average of six subjects). The coefficient of variation (i.e., standard deviation divided by the mean) was always less than 0.08 and usually lower than 0.03. The presence of a speech band (“slit”) at each of three center frequencies (750, 1500 and 3000 Hz) is indicated by either “–” (no low-pass modulation filtering) or “•” (low-pass modulation filtered). The low-pass modulation filter cutoff varied between 3 and 24 Hz. 99% of the consonants were correctly identified in the absence of spectral and modulation filtering (i.e., unprocessed stimuli).

Slit Frequency			Low-Pass Modulation Filtering				
750	1500	3000	AllPass	< 24 Hz	< 12 Hz	< 6 Hz	< 3Hz
•			38.4	32.8	27.5	21.5	18.2
	•		40.2	35.9	29.0	19.7	16.2
		•	39.6	31.3	29.0	21.5	16.7
•	•		67.6	62.1	55.6	41.7	26.3
	•	•	77.1	76.4	71.7	56.8	34.8
•		•	74.6	73.2	63.6	46.0	31.6
•	•	•	88.4	87.9	81.1	64.1	42.9
•	–			64.8	59.1	57.1	50.0
•		–		69.7	55.3	50.3	47.0
	•	–		76.0	71.5	67.2	61.4
–	•			64.3	60.9	57.8	45.5
–		•		71.5	59.6	56.6	51.3
	–	•		77.7	73.0	68.4	60.4
•	–	–		85.9	84.6	83.6	79.0
–	–	•		87.1	85.4	80.1	76.0
–	•	–		78.3	79.5	74.5	71.5
•	–	•		86.6	82.3	75.8	65.9
•	•	–		82.8	78.8	77.3	66.7
–	•	•		87.9	84.1	75.0	61.4

combination with one or two other slits and low-pass modulation -filtered using the “Modulation Toolbox” [7]. The low-pass cutoff frequency of modulation filtering ranged between 3 Hz and 24 Hz in octave steps. Each slit was also presented without any modulation filtering. The stimulus was preceded by a short, unfiltered carrier phrase “På pladsen mellem hytten og...” and contained one of eleven consonants, [p, t, k, b, d, g, m, n, f, s, v], followed by one of three vowels, [i, a, u]. In the current study, the impact of vocalic environment on consonant identification (and phonetic-feature decoding) was not considered due to limitations of space and time. Each token concluded with the liquid segment [l] (e.g., “talle,” “tulle,” “tille”). The full set of stimulus conditions is listed in Table 1. Table 2 lists the phonetic features associated with each consonant.

The material was spoken by two talkers (one male, one female), and presented diotically over Sennheiser HD-580 headphones at an average

Table 2: 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. Code: + (presence); – (absence); S (stop); F (fricative); N (nasal); A (anterior); C (central); P (posterior)

Consonant	Voicing	Manner	Place
p	–	S	A
t	–	S	C
k	–	S	P
b	+	S	A
d	+	S	C
g	+	S	P
s	–	F	C
f	–	F	A
v	+	F	A
m	+	N	A
n	+	N	C

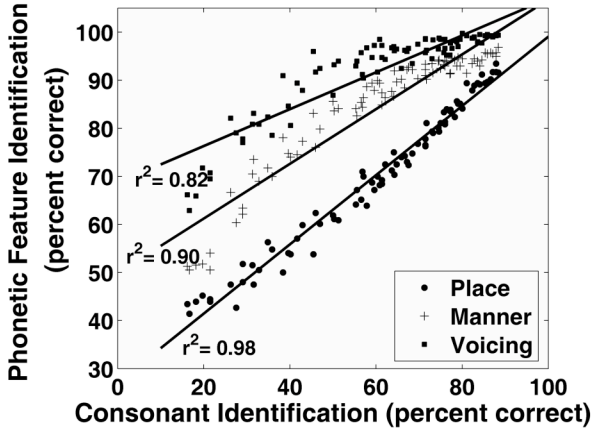
sound pressure level of 65 dB. The subject was seated in a double-walled sound booth, and was asked to identify the initial consonant of each stimulus. No feedback was provided. Six individuals (3 males, 3 females) between the ages of 21 and 28 years old participated in the study. All reported normal hearing and no history of auditory pathology.

3. DATA ANALYSIS AND RESULTS

The data were analyzed in several ways. First, consonant identification accuracy progressively declines with decreasing low-pass modulation-frequency cutoff (Table 1). Second, the number of slits affects consonant recognition accuracy. Consonant identification was also scored in terms of how well a consonant’s phonetic features – voicing, manner and place of articulation – were decoded. When a consonant is correctly identified, its constituent phonetic features are (by definition) also decoded accurately. However, when a consonant is incorrectly recognized, it is rare that all of its constituent phonetic features are incorrectly decoded; one or two features are usually decoded accurately (more often voicing and manner than place – see below).

Consonant perception is usually studied in terms of accuracy for individual phonetic segments. Because consonants are phonetically related to each other, scoring *only* in terms of the proportion of consonants correct may obscure patterns associated with cross-spectral integration and modulation processing. Confusion matrices of

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



consonant-identification error patterns provide a straightforward means of delineating how much information associated with constituent phonetics features is transmitted. To compute the “true” amount of information associated with each consonant, a bias-neutral metric (such as d' used in signal detection theory) is desirable. The information transmitted [6] was computed by partitioning the 11 consonants into three (overlapping) groups of voicing, articulatory manner and place of articulation. Voicing is a binary distinction, whereas manner and place encompass three class distinctions (Table 2).

Confusion matrices were computed for each phonetic feature. In essence, each phonetic feature is treated as a *quasi*-independent information channel (because of space limitations, phonetic-feature interaction lies outside the scope of this paper). Although a consonant may be identified incorrectly, there may be information pertaining to its constituent phonetic properties that is correctly decoded. Information associated with voicing and manner of articulation is generally decoded accurately even when the consonant is not identified correctly (Figure 1). In contrast, place of articulation is rarely decoded correctly when the consonant is misidentified. Such analyses suggest that consonant identification depends *largely* on decoding the place of articulation dimension correctly.

In order to compute the amount of information associated with a specific feature and stimulus condition, it is necessary to calculate the covariance between a specific stimulus and response

category (as a means of neutralizing the effect of response bias). The information associated with voicing, manner and place is computed as follows:

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

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.

When the data are plotted in terms of the amount of information transmitted, interesting patterns emerge (Table 3). Information combines differently across the audio spectrum for each phonetic feature. In the absence of low-pass modulation filtering, both voicing and manner information combine linearly for two-slit signals. For three-slit stimuli, voicing information saturates (i.e., contains the same amount of information as the two-slit signals), while manner information increases somewhat (i.e., exhibits some compression). In contrast, place of articulation combines *synergistically* (two or three slits contain far more information than linear summation would predict). There is substantially greater than linear summation across slits for virtually all conditions. The amount of place of articulation information transmitted within any single slit is small (substantially less than manner or voicing). This suggests that place depends largely on cross-spectral integration. It is also the feature most important for decoding consonants. From such patterns we conclude that cross-spectral integration is particularly important for speech robustness (given the importance of consonant decoding for spoken language comprehension).

There is also a progressive decline in place and manner information transmitted with low-pass filtering of the modulation spectrum. This decline is particularly pronounced above 6 Hz for single-slit stimuli. In contrast, voicing information is most sensitive to modulation filtering *below* 6 Hz. When two or three slits are combined, phonetic feature information is relatively unaffected by modulation filtering as long as modulation frequencies greater than 6 Hz are preserved. When the modulation spectrum is filtered below 6 Hz, cross-spectral integration becomes extremely important for decoding all features (i.e., voicing, place and manner of articulation).

Table 3: The amount of transmitted information (as specified in Equation 1) computed for each phonetic feature (place, manner, voicing) in conditions where each slit undergoes the same amount of low-pass modulation filtering). The signals contain 1, 2 or 3 spectral slits (whose center frequencies are indicated). Bold cells indicate conditions in which lowering the low-pass modulation cut-off frequency by one step (e.g. from 24 to 12 Hz) resulted in a significant decline ($\geq 25\%$) of transmitted information. Cells marked by a single asterisk (*) indicate where cross-spectral integration of transmitted information is more than 50% greater than predicted on the basis of linear summation. Cells marked by a double asterisk (**) indicate where cross-spectral integration is more than 200% greater than predicted on the basis of linear summation.

		Slit Center Frequencies						
		750	1500	3000	750	1500	750	1500
Low Pass Modulation Filtering		750	1500	3000	1500	3000	3000	3000
	P L A C E	All Pass	0.14	0.10	0.09	0.41	**0.72	*0.62
<24 Hz		0.09	0.13	0.07	*0.40	**0.74	**0.59	**1.12
<12 Hz		0.03	0.05	0.06	**0.27	**0.65	**0.38	**0.94
<6 Hz		0.02	0.01	0.02	**0.11	**0.37	**0.21	**0.47
<3 Hz		0.02	0.01	0.02	*0.05	**0.19	*0.07	**0.27
M A N N E R	All Pass	0.58	0.45	0.42	1.04	0.96	1.10	1.24
	<24 Hz	0.42	0.36	0.31	0.85	1.10	1.00	1.18
	<12 Hz	0.22	0.22	0.16	*0.80	*0.98	*0.87	*1.04
	<6 Hz	0.10	0.09	0.07	**0.59	**0.72	**0.55	**0.84
	<3 Hz	0.11	0.06	0.04	*0.27	**0.32	*0.41	*0.51
V O I C E	All Pass	0.55	0.30	0.39	0.79	0.72	0.90	0.94
	<24 Hz	0.31	0.25	0.30	0.68	0.72	0.87	0.95
	<12 Hz	0.27	0.23	0.22	0.66	*0.77	*0.79	0.94
	<6 Hz	0.11	0.14	0.12	*0.51	*0.56	*0.59	*0.81
	<3 Hz	0.07	0.07	0.04	**0.33	*0.33	**0.37	*0.48

4. CONCLUSIONS

Conventional methods of estimating the contribution made by different parts of the audio (frequency) spectrum and the modulation (temporal) spectrum generally 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 information-theoretic analysis used in this study can be used to delineate precisely which parts of the speech signal are of greatest importance for decoding phonetic features associated with intelligibility and comprehension in a way that is difficult to achieve by analyzing consonant identification scores alone.

We conclude from the data illustrated in Figure 1 and Table 3 that cross-spectral integration of modulation patterns (particularly those associated with place-of-articulation information) is crucial for accurate decoding of spoken language. Moreover, cross-spectral integration is likely to hold the key for improving intelligibility in acoustically challenging environments and is particularly important for ameliorating the deficits of the hearing impaired.

5. REFERENCES

- [1] 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.
- [2] Drullman, R., Festen, J.M., Plomp, R. 1994. Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.* 95, 2670-2680.
- [3] Greenberg, S., Arai, T. 2004. What are the essential cues for understanding spoken language? *IEICE Trans. Inf. & Syst.* E87-D, 1059-1070.
- [4] Greenberg, S., Arai, T., Silipo, R. 1998. Speech intelligibility derived from exceedingly sparse spectral information. *Proc. 5th Int. Conf. Spoken Lang. Proc.*, 74-77.
- [5] Houtgast, T., Steeneken, H.J.M. 1985. A review of the MTF-concept in room acoustics. *J. Acoust. Soc. Am.* 77, 1069-1077.
- [6] Miller, G.A., Nicely, P.E. 1955. An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.* 27, 338-352.
- [7] Schimmel S., Atlas, L. 2005. Coherent envelope detection for modulation filtering of speech. *Proc. Int. Conf. Audio, Speech & Signal Proc. (ICASSP)*, 221-224.
- [8] Silipo, R., Greenberg, S., Arai, T. 1999. Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations, *Proc. 6th European Conf. Speech Comm. Tech. (Eurospeech-99)*, 2687-2690.