Spectro-temporal processing of speech – An information-theoretic framework

Thomas U. Christiansen[1], Torsten Dau[1], and Steven Greenberg[1,2]

[1] Center for Applied Hearing Research, Ørsted•DTU, Acoustic Technology, Technical University of Denmark, Ørsteds Plads, bldg. 352, DK-2800 Kgs. Lyngby, Denmark {tuc, tda}@oersted.dtu.dk

[2] Silicon Speech, 46 Oxford Drive, Santa Venetia, CA 94903, USA, steveng@savant-garde.net

1 Introduction

Which acoustic cues are important for understanding spoken language? Traditionally, the speech signal is described mainly in spectral terms (i.e., the distribution of energy across the acoustic frequency axis). In contrast, temporal properties are often ignored. However, there is mounting evidence that low-frequency energy modulations play a crucial role, particularly those below 16 Hz (e.g., Christiansen and Greenberg 2005; Drullman, Festen and Plomp 1994; Greenberg and Arai 2004; Houtgast and Steeneken 1985). Modulations higher than 16 Hz may also contribute under certain conditions (Apoux and Bacon 2004; Christiansen and Greenberg 2005; Greenberg and Arai 2004; Silipo, Greenberg and Arai 1999). Currently lacking is a detailed understanding of how amplitude-modulation cues are combined across the acoustic frequency spectrum, as well as how spectral and temporal information *interact*. Such knowledge could 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 systematic filtering of the modulation spectrum 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. For this reason, amplitude modulations associated with each of three discrete spectral regions were low-pass filtered, 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 feature of voicing (e.g., differentiating the consonants, [p, t, k] from [b, d, g]), articulatory manner (e.g., distinguishing [b] from [m], [d] from [n]) and place of articulation (e.g., distinguishing [p] from [t]. and [k]) can be used to assess the contribution of each audio-frequency channel and modulation-frequency region to consonant recognition 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 can be discerned when presented alone and in combination with other spectral bands (see Christiansen and Greenberg 2005 for details).

Stimuli were Danish monosyllabic words and nonsense syllables recorded in a sound-insulated environment at Aalborg University. 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. 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 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 and low-pass modulation filtered using the "Modulation Toolbox" (Atlas and Thompson 2004). 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]. Each token concluded with the liquid segment [l] (e.g., "talle," "tulle," "tille"). The full set of stimulus conditions is listed in Table 1.

The material was spoken by one of two talkers (one male, one female), and presented diotically over Sennheiser HD-580 headphones at a sound pressure level of 65 dB to the subject, who was seated in a double-walled sound booth. The subject‘s task was 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 participated in the study. All reported normal hearing and no history of auditory pathology.

3 Data analysis and results

The data were analyzed in a variety of ways. Consonant identification accuracy declines progressively as a function of low-pass modulation-frequency cutoff (Table 1). The number of slits also 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.

Slit Frequency			**Low-Pass Modulation Filtering**				
750	**1500**	**3000**	**All Pass**	**< 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
			39.7	31.4	29.0	21.5	16.7
			62.9	61.6	55.6	41.7	26.3
			73.5	75.0	71.7	56.8	34.9
			69.2	71.0	63.6	46.0	31.6
			88.4	87.9	81.1	64.1	42.9
	•			65.4	59.1	57.1	50.0
		•		63.1	55.3	50.3	47.0
		•		74.5	71.5	67.2	61.4
•				61.9	60.9	57.8	45.5
•				63.1	59.6	56.6	51.3
	•			75.5	73.0	68.4	60.4
	•	•		85.9	84.6	83.6	79.0
•	•			87.1	85.4	80.1	76.0
•		•		78.3	79.6	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

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).

Consonant perception is usually studied in terms of accuracy for individual 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 consonantal identification error patterns provide a straightforward means of delineating how much information associated with constituent phonetics features is transmitted. In order to compute the "true" amount of information associated with each consonant, a bias-neutral metric (such as *d´* used in signal detection theory) is required. To compute the amount of information transmitted (Miller and Nicely, 1955), the eleven consonants were partitioned 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 for the Danish consonants used in this study.

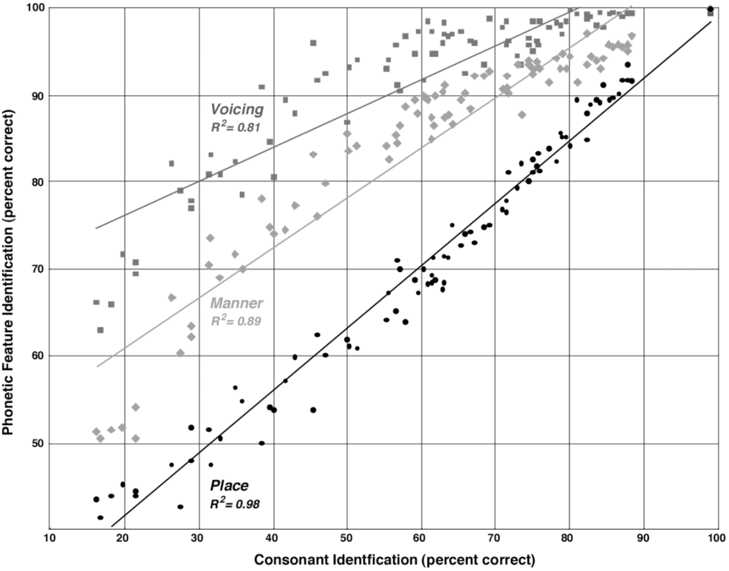

Figure 1. Consonant identification accuracy as a function of phonetic feature classification for the same conditions and listeners shown in Table 1. The correlation coefficient (R^2) is shown for each phonetic feature.

Confusion matrices were computed for each phonetic feature (see Christiansen and Greenberg 2005 for details). In essence, each phonetic feature is treated as a quasi-independent information channel. Although a consonant may be identified incorrectly, there may be information pertaining to its constituent phonetic properties that is correctly decoded. Information pertaining to 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 accurately when the consonant is incorrectly identified. Such analyses demonstrate that consonant identification depends largely on decoding the place-of-articulation feature correctly.

In order to compute the amount of information associated with a specific feature and stimulus condition, it is necessary to calculate the co-variance 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:

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

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.

When the data are plotted in terms of the amount of information transmitted, interesting patterns emerge (Table 2). 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 (contains the same amount of information as the two-slit signals), while manner information is slightly compressed. In contrast, place of articulation combines synergistically (two or three slits contain far more information than linear summation would predict). Cross-spectral integration is particularly important for speech robustness. Place of articulation is the phonetic feature that depends most on cross-spectral integration. There is substantially greater than linear summation across slits for virtually all conditions. The amount of information transmitted within any single slit is small (substantially less than manner or voicing). Hence, place information requires a broad span of the speech spectrum to be decoded accurately. Place of articulation is also the feature most closely associated with the ability to accurately decode consonant identity (Figure 1).

There is a progressive decline in place and manner information transmitted with low-pass filtering of the modulation spectrum, particularly 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 three phonetic features (voicing, place and manner of articulation). Amplitude-modulation cues present in separate audio frequency regions largely compensate for the low-pass filtering of the modulation spectrum.

		Slit Center Frequencies						
	Low Pass Modulation Filtering	**750**	**1500**	**3000**	**750** **1500**	**1500** **3000**	**750** **3000**	**750** **1500** **3000**
PLACE	**All Pass**	0.14	0.10	0.10	0.32	**0.67	*0.53	**1.03
	<24 Hz	**0.09**	0.13	**0.07**	*0.38	**0.69	**0.53	**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**
MANNER	**All Pass**	0.58	0.45	0.42	0.97	0.81	1.04	1.24
	<24 Hz	**0.42**	**0.36**	0.31	0.86	1.07	0.97	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**
VOICE	**All Pass**	0.56	0.3	0.39	0.76	0.65	0.90	0.94
	<24 Hz	**0.31**	0.25	0.30	0.70	0.71	0.84	0.95
	<12 Hz	0.27	0.23	0.22	0.67	*0.77	*0.80	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**

Table 2. 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 the low-pass modulation filtering result in a significant decline (≥ 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.

4 Conclusions and significance

Conventional methods of estimating the contribution made by different parts of the audio spectrum and the modulation spectrum 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. We conclude that:

(1) Place of articulation information is crucial for accurate consonant recognition.

(2) Accurate decoding of place-of-articulation information requires broadband cross-spectral integration.

(3) Place of articulation information is associated most closely with the modulation spectrum between 6 and 24 Hz. Hence, consonant decoding requires cross-spectral integration of the modulation spectrum above 6 Hz.

(4) Voicing is mainly associated with the modulation spectrum *below* 6 Hz.

(5) Manner of articulation is associated with the modulation spectrum between 3 and 24 Hz for single-band stimuli. For signals containing two or more slits, the modulation spectrum below 6 Hz may be particularly important, especially in noisy and reverberant conditions.

(6) Cross-spectral integration of modulation patterns is crucial for accurate decoding of spoken language, and therefore holds the key for improving intelligibility in acoustically challenging environments.

References

Apoux, F. and 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.

Atlas, L., Li, Q. and Thompson, J. (2004) Homomophic modulation spectra. *Proc. Int. Conf. Audio, Speech & Signal Proc. (ICASSP),* pp. 761-764.

Christiansen, T. U. and Greenberg, S. (2005) Frequency selective filtering of the modulation spectrum and its impact on consonant identification. In: A. Rasmussen and T. Poulsen (Eds.), *Twenty First Danavox Symposium*, pp. 585-599.

Drullman, R., Festen, J.M. and Plomp, R. (1994) Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.* 95, 2670-2680.

Greenberg, S. and Arai, T. (2004) What are the essential cues for understanding spoken language? *IEICE Trans. Inf. & Syst.* E87-D, 1059-1070.

Greenberg, S., Arai, T. and Silipo, R. (1998) Speech intelligibility derived from exceedingly sparse spectral information. *Proc. 5th Int. Conf. Spoken Lang. Proc.*, pp. 74-77.

Houtgast, T. and Steeneken, H.J.M. (1985). A review of the MTF-concept in room acoustics. *J. Acoust. Soc. Am.* 77, 1069-1077.

Miller, G.A. and Nicely, P. E. (1955) An analysis of perceptual confusions among some English consonants. *J. Acoustic. Soc. Am.* 27, 338-352.

Silipo, R., Greenberg, S. and Arai, T. (1999) Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations, *Proc. 6th European Conf. Speech Comm. Tech. (Eurospeech-99)*, pp. 2687-2690.