Linguistic scene analysis and the importance of synergy

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

[1] Silicon Speech, Santa Venetia, CA 94903, USA

2 Centre for Applied Hearing Research, Ørsted•DTU, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

This chapter explores the possibility that speech is decoded using cross-spectral and cross-modal integration strategies that are *inherently* synergistic. Combining information from separate spectral channels or across modalities may result in far greater intelligibility and phonetic recognition than predicted by linear-integration models. This is because decoding speech relies on multi-tier processing strategies that are opportunistic and idiosyncratic. Models incorporating synergistic integration are more likely to predict linguistic comprehension than conventional, linear approaches, particularly in challenging listening conditions.

LINGUISTIC SCENE ANALYSIS

Linguistic scene analysis (LSA) is the process by which the listener analyzes and *interprets* the acoustic and visual sensory streams in the process of understanding a talker's message. *Interpretation is key to speech understanding.* This is because the talker communicates within a specific information framework associated with a behavioural context. Without context, the sensory streams associated with language are difficult to decode. Context provides not just the grammatical and semantic framework but also the behavioural framework (a.k.a. "pragmatics"). The sounds of spoken language are but one source of information with which to decode the message. The same sequence of segments (e.g., [y eh s] "yes") can have very different meanings depending on what comes before or after. Conversely, an indecipherable babble can convey an unambiguous meaning when embedded in the appropriate context. Also important is the listener's *internal state,* which relates to such extra-linguistic dimensions as memory, personality and intention.

Only some of the variables germane to speech communication are *observable.* Many are "hidden" and can only be deduced through clever, intricate experimentation (if at all) (Greenberg, 2007). Moreover, the brain rarely acts like a *linear* integrator of sensory streams. This is why visual cues can boost intelligibility far more in noisy and reverberant conditions than would be predicted when presented alone. The brain specializes in combining cues from disparate sources to derive very specific information difficult to derive from a *single* source.

Linear models are unlikely to predict speech intelligibility under conditions of greatest interest, namely the "real world." For real-world modelling, a more sophisticated approach is required, one that focuses on *synergy* rather than linear integration. In this chapter, one aspect of LSA – phonetic decoding and its potential relation to prosody – is examined as an example of the analyses required to gain insight into how the brain

Auditory signal processing in hearing-impaired listeners. 1st International Symposium on Auditory and Audiological Research (ISAAR 2007). T. Dau, J. M. Buchholz, J. M. Harte, T. U. Christiansen (Eds.).
ISBN: 87-990013-1-4. Print: Centertryk A/S.

goes from sound to meaning.

WHAT UNDERLIES THE CONTEXT EFFECT?

Intelligibility of an acoustic signal varies depending on its "linear" (i.e., left to right, time-flow) context. Within a grammatically (and semantically) well-formed utterance, intelligibility is *much* greater for words embedded within such a context relative to that associated with the same *acoustic signal* presented in isolation (Fig. 1). What is responsible for this enormous gain in intelligibility?

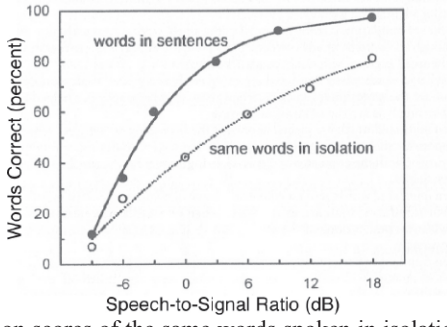

Fig. 1: Identification scores of the same words spoken in isolation and in sentences as a function of speech-to-noise ratio [adapted from Miller, Heise and Lichten, (1951) by Plomp (2002), p. 106 (axes labels redrawn for legibility)].

The traditional "answer" is "context." But what does context refer to? Usually, it refers to the grammatical and semantic structure in which individual words are embedded. For reasons poorly understood, a *sequence* of words is easier to understand than the same words presented in *isolation.* This phenomenon is illustrated in Fig. 2. Between three and five words spoken in sequence are required for a listener to achieve more than 80% intelligibility. Clearly, listeners don't decode the speech signal one word at a time. If they did, the data in Fig. 2 would be fictitious. What is there about lexical sequencing that makes it easier to decode individual words (spoken in context)?

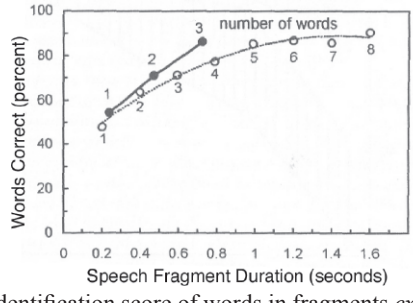

Fig. 2: Average identification score of words in fragments *excised* from read text (solid symbols) and conversational speech (open symbols) as a function of fragment duration [based on data from Pickett and Pollack (1963) and Pollack and Pickett (1963) and adapted from Plomp, 2002, p. 107 (axes labels redrawn for legibility)].

Fig. 3 suggests that the answer is rather complicated. Grammatical structure *does* appear to enhance intelligibility when the words are monosyllabic. *However,* there is relatively little (if any) gain due to grammatical context when the words contain two or more syllables. Why?

What distinguishes words of a single syllable from those composed of polysyllables? There are many possibilities, of course. Among the most prominent and consistent is prosody. Prosody refers to a linguistic attribute associated with syllable sequences. A syllable is either "stressed" or not. Stressed syllables tend to be longer and louder than unstressed ones. They are perceptually more "prominent" and therefore tend to stand out from their unstressed counterparts. It is rare for all syllables in an utterance to be *exclusively* stressed or unstressed (unless the utterance consists of a single syllable). Normally, there is a patterned variation in stress among successive syllables that imparts a certain rhythmic structure (Greenberg, 2006). Utterances spoken in an inappropriate rhythm are usually deemed odd or foreign, and are often more difficult to understand even if the phonetic constituents are all present.

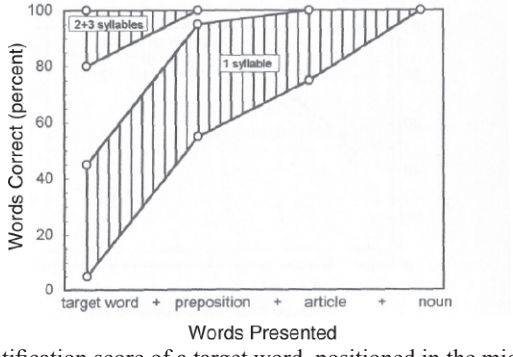

Fig. 3: Average identification score of a target word, positioned in the middle of a sentence, as a function of the number of words presented following the target word (based on data from Grosjean, 1985) [Adapted from Plomp, 2002, p. 108].

A word containing two or more syllables is far more likely to have a pronounced rhythm than its monosyllabic counterpart. Monosyllabic words may participate in a prosodic pattern when combined with other words. This lexical sequence is commonly referred to as a linguistic phrase. In Fig. 3, the sequence "Target Word + Preposition + Article + Noun" would constitute such a phrasal unit. Two monosyllabic words may be as intelligible as a polysyllabic word under certain circumstances. It is tempting to speculate that the variability in intelligibility observed among monosyllabic words in Fig. 3 is attributable to variation in prosody. For the present, this hypothesis remains speculative.

How does prosody facilitate speech understanding? We don't really know. However, there is some intriguing evidence that at least *part* of the context effect is attributable to prosodic patterning, a possibility we consider in greater detail later in this chapter. The beneficial impact of context probably reflects synergistic processes, though it is usually not treated in this way. We next consider another form of synergy that impacts intelligibility.

INTELLIGIBILITY AND ITS DISCONTENTS

The Articulation Index (AI – French and Steinberg, 1947; ANSI, 1997) and Speech Transmission Index (STI – Houtgast and Steeneken, 1985) are two popular methods for estimating the intelligibility of acoustic speech signals. Under certain conditions, both metrics estimate intelligibility reasonably well. The conceptual basis of the AI and STI are similar. They both assume that acoustic-frequency channels with the highest signal-to-noise ratio (SNR) contribute most to intelligibility, and that intelligibility reflects some form of quasi-linear integration across the acoustic frequency spectrum.

The problem with these metrics concerns the conditions in which neither is able to accurately predict intelligibility. If 80% of the acoustic spectrum is discarded, with minimal intelligibility associated with the remaining individual spectral channels (when presented alone), most listeners can still understand spoken sentences extremely well (Greenberg *et al.*, 1998). The key is to retain the "right" 20% of spectral channels. As long as the acoustic spectrum is sampled in a uniform way, most of it is dispensable (under optimal listening conditions). Others have reached similar conclusions (e.g., Müsch and Buus, 2001). The intelligibility of "sparse spectral speech" is at odds with linear integration of acoustic information.

An even greater challenge arises when visual speech information is considered. Intelligibility of the visual stream presented alone is usually quite low – typically 10% or less. When this sensory signal is combined with sparse-spectral speech (in this instance, two, one-third-octave slits, one centred at 330 Hz the other at 5400 Hz) something very interesting occurs. Intelligibility of the two-slit acoustic signal is ca. 20%. When the visual and acoustic signals are combined, intelligibility jumps to 63%, which is about double what is predicted by the product-of-errors heuristic used in the AI (Fig. 4). Interestingly, the ANSI standards version of the AI, known as the Speech Intelligibility Index (SII – ANSI, 1997), specifically excludes conditions where acoustic and visual signals are combined. Unfortunately (for the AI and STI metrics), most speech communication is conducted face-to-face, even in this age dominated by mobile phones.

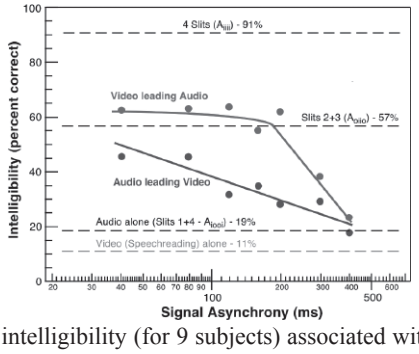

Fig. 4: Average intelligibility (for 9 subjects) associated with audio-visual speech recognition as a function of bi-modal signal asynchrony. The audio-leading-video conditions are marked in blue, the video-leading-audio conditions shown in red. Baseline audio-only conditions are marked in black, dashed lines, and the video-alone condition is shown in orange. From Grant and Greenberg (2001).

The SII is unlikely to be appropriately revised to accommodate visual information. This is because the linear-integration approach assumes that the incoming sensory streams are processed symmetrically in time. A temporal jitter in one direction should be roughly equivalent to time jitter in the opposite direction. For acoustic signals, this appears to be (approximately) the case (Silipo *et al.*, 1999). However, for audio-visual speech there is a pronounced asymmetry. Delaying the visual signal relative to the audio results in a precipitous decline in intelligibility (Fig. 4). However, when the visual signal leads the audio, there is virtually no impact on intelligibility unless the temporal disparity exceeds 200 ms (Fig. 4). Such intelligibility effects are inconsistent with linear integration models of speech perception.

Why does the visual stream enhance the audio to such a degree? And what accounts for the temporal asymmetry in combining the sensory streams? To answer such questions (and to gain deeper insight into how the brain decodes spoken language), a finer-grained linguistic level than word intelligibility is required, one that focuses on phonetic (articulatory-acoustic) features.

AN ATOMISTIC PERSPECTIVE OF SPEECH

It is well known that utterances are composed of prosodic phrases, which are sequences of words. A word is composed of one or more syllables, each containing a certain number of phonetic segments. A segment can be broken down into a constellation of features derived from articulatory gestures. The three principle articulatory-feature dimensions are *voicing* (reflecting the vibration of the laryngeal vocal folds), *manner of articulation* (associated with the mode of production and the way in which air passes through the vocal tract) and *place of articulation* (reflecting the vocal tract locus of maximum occlusion) (Greenberg, 2006). In principle, any segment can be uniquely specified by its articulatory feature specification (some segments require additional features to be uniquely distinctive).

Linguistically, these three feature dimensions possess distinctive properties. Voicing and manner are closely associated with the syllable's energy contour. On average, approximately 80% of the speech signal is voiced. The unvoiced parts are always on the flanks of syllables (except in those rare instances where the entire syllable is unvoiced) and lie in the low-energy part of the contour. In this sense, voicing reflects the build-up (and decline) of energy across the syllable. The intonation (fundamental frequency) contour is associated with the voiced parts of the speech signal (though perceptually, the contour appears to continue through the unvoiced portions). The harmonic structure associated with voicing also serves to shield the signal's message from acoustic interference (e.g., speech babble and reverberation). Although voicing can serve to distinguish segments and words (e.g., "ba<u>t</u>" vs. "ba<u>d</u>" or "<u>b</u>at" vs. "<u>p</u>at"), in the everyday world of spontaneous discourse, it rarely does so. Usually, semantic context narrows the phonetic options to a point where the voicing distinction is superfluous. In this sense, voicing is the least lexically distinctive feature dimension (the importance of this point will be apparent later in this chapter).

Manner of articulation is also associated with the syllable's energy contour. Certain manner classes, such as the stops and fricatives, typically occur in either the syllable onset or coda (i.e., end) where the energy level is relatively low. Other classes, such as vowels, glides and liquids contain far more energy and often occupy the syllable's centre (nucleus). The order in which phonetic segments occur within the syllable ("phonotactics") is governed by an "energy arc" principle in which the low-energy manner classes flank those of higher energy (Greenberg, 2006). The energy arc is important for packaging the acoustic signal in a way that the auditory system (and other brain regions) can "digest." The low-frequency portion (2–6 Hz) of the speech signal's modulation spectrum reflects this quasi-periodic energy fluctuation. Manner is more important for distinguishing words than voicing, but not by much. Certain manner classes (of roughly comparable energy), such as stops and fricatives, can substitute for each other under many speaking conditions (e.g., a stop can often be articulated as a fricative without significant impact on intelligibility. Also, it is rare for two segments of the same manner class to be adjacent within the same syllable (the exceptions are morphologically significant – "look" [l uh k] "looked" [l uh k t], consistent with the notion that the sequence of manner classes is "designed" to guide the energy into an arc-like contour.

In contrast to manner and voicing, whose specific identity may not matter all that much for intelligibility, place of articulation is usually crucial for lexical discrimination. Unlike voicing and manner, which are relatively coarse energy features, place of articulation encodes detailed spectral patterns that are largely independent of energy level. These spectral patterns are associated with distinct locations in the vocal tract where the articulators come in close proximity (i.e., maximum occlusion). The locus of airflow constriction leaves an "acoustic fingerprint" in terms of spectral maxima. In English, there are technically ten distinct loci of constriction. However, virtually all of these are closely linked to a specific manner of articulation. In actual practice, it is rare for a manner class to have more than three distinct places of articulation (front, back, in-between). Therefore, if the manner is known, it greatly reduces the complexity of identifying place (Chang *et al.*, 2005). In this sense, decoding place of articulation depends on manner, a point we'll consider later in this chapter.

The visual cues for speech are thought to be most closely associated with place of articulation. Most of the phonetic confusions observed in audio-visual tests of nonsense-syllable identification are place errors. Approximately 94% of *phonetic* information provided by the visible articulators is place-related (Grant and Walden, 1996). Is this why visual speech cues are such a powerful adjunct of the acoustic signal in noisy backgrounds (and for the hearing impaired)? Before examining this issue, let's first consider how place of articulation (and other phonetic feature dimensions) are decoded in the *acoustic* signal.

ERROR PATTERNS OF CONSONANT CONFUSIONS

Over fifty years ago, Miller and Nicely (1955) developed a method for ascertaining the contribution of each phonetic-feature dimension to consonant identification. Instead

of merely computing the percent of consonants correctly reported, they examined the patterns of errors as a means of determining whether mistakes are uniformly distributed (or not). In the conditions observed (variable amounts of background noise), the vast majority of errors were associated with place of articulation (this can easily be computed by determining which consonants are confused with others – place errors occur when confusing [p] with [t] or [k], [d] with [b] or [g], [m] with [n], and so on). Details of the computational procedure are described in Christiansen and Greenberg (2005) and Christiansen *et al.* (2007).

The Miller and Nicely confusion paradigm can be adapted to other material, in this instance, Danish consonants (Christiansen and Greenberg, 2005; Christiansen *et al.*, 2007). Stimuli were Danish monosyllabic words and nonsense syllables. The acoustic frequency spectrum was partitioned into three separate channels ("slits"), each three-quarters of an octave wide. The lowest slit was centred 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 others. Each slit, when presented alone, resulted in ca. 40% consonant recognition. Three slits presented concurrently were identified correctly ca. 90% of the time. This 50% dynamic range in consonant identification allows us to observe the process by which the auditory system and brain decode phonetic segments and features.

The relation between consonant identification and phonetic-feature decoding is shown in Fig. 5. Note that voicing and manner decoding is relatively accurate (and well above chance level of performance) even when consonant identification is poor (ca. 40% correct). This pattern is significant because it implies that certain phonetic properties of the speech signal are accurately decoded even under highly degraded conditions. We'll return to this point later in this chapter.

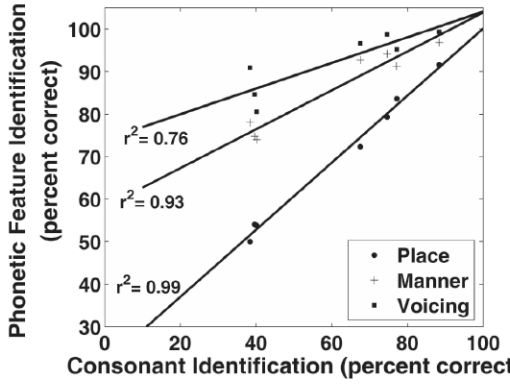

Fig. 5: Voicing, manner and place of articulation decoding precision as a function of overall consonant identification accuracy. For each phonetic-feature dimension a best-fit linear regression has been computed (r2). Plots are based on data from six Danish listeners. [adapted from Christiansen and Greenberg (2008)].

The other important point to note is the near-perfect correlation between place-of-

articulation decoding and consonant identification (r2 = 0.99). Such high degrees of correlation are rarely observed in experimental data and are usually indicative of a strong pattern. In this instance, they suggest that consonant identification is crucially dependent on decoding place cues correctly.

Another way of stating these data patterns is as follows: all three phonetic-feature dimensions (voicing, manner and place) need to be correctly decoded in order for a consonant to be identified correctly. However, decoding manner and voicing information is less crucial than place in this process. When a consonant is incorrectly identified, it is common for both manner and voicing to be correctly decoded (as deduced from the confusion patterns). Not so for place. It is extremely rare that place information is decoded correctly when a consonant is reported incorrectly. In this sense, place of articulation appears to underlie the ability to recognize and identify specific consonants.

DECODING PLACE OF ARTICLATION USING CROSS-SPECTRAL SYNERGY

It is possible to deduce how consonants are processed in the auditory system by transforming the error patterns into an information-theoretic metric. To compute the amount of information transmitted, the eleven Danish consonants were partitioned into three (overlapping) groups of voicing, manner and place of articulation as shown in Table I. As a means of neutralizing the effect of response bias, it is necessary to compute the amount of information (in bits) associated with a specific phonetic feature and stimulus condition by calculating the co-variance between a specific stimulus and response category. The information associated with voicing, manner and place is computed as follows (based on Miller and Nicely, 1955):

$$T(x,y) = -\sum_{i,j} p_{ij} \log \frac{p_i p_j}{p_{ij}} \qquad \text{(Eq. 1)}$$

where *T(x,y)* refers to the number of bits per feature transmitted from *x* to *y*, 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 (Fig. 6). Information combines differently across the frequency spectrum for each phonetic feature. Both voicing and manner information combine quasi-linearly for two-slit signals. For three-slit signals, voicing information contains the same amount of information as the two-slit signals, while manner information is slightly compressed. In contrast, place of articulation combines synergistically (i.e., two or three slits contain *far* more information than linear summation would predict. Place of articulation is the phonetic feature dimenson 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 substantially less than manner or voicing. The implication is that place information requires a broad span of the speech spectrum to be decoded correctly.

Segment	Voicing	Manner of Articulation	Place of Articulation
[p]	Voiceless	Stop	Anterior
[t]	Voiceless	Stop	Medial
[k]	Voiceless	Stop	Posterior
[b]	Voiced	Stop	Anterior
[d]	Voiced	Stop	Medial
[g]	Voiced	Stop	Posterior
[s]	Voiceless	Fricative	Medial
[f]	Voiceless	Fricative	Anterior
[v]	Voiced	Fricative	Anterior
[n]	Voiced	Nasal	Medial
[m]	Voiced	Nasal	Anterior

Table 1: The phonetic features associated with the 11 Danish consonants used in the study. Voicing is a binary feature, while Manner and Place are ternary-valued features.

The difference in cross-spectral integration among the phonetic features is highlighted in Fig. 7. The cross-spectral integration quotient (XS IQ) is the ratio of the observed information transmission for a given multi-band condition and the sum of information associated with contributing individual bands. If the integration is linear, the XS IQ will be close to one for the two-slit conditions. This is the case for voicing, manner and consonants.

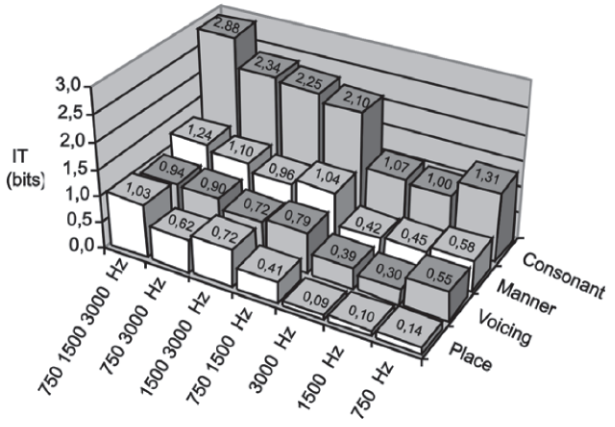

Fig. 6: Information transmitted associated with consonant identification as well as decoding of the phonetic-feature dimensions of voicing, manner and place of articulation. [adapted from Christiansen and Greenberg (2008)].

However, the XS IQ for place is much higher – between 1.72 and 3.65, suggesting that place information is integrated in a highly non-linear way. The non-linearity is

expansive, meaning that the amount of information associated with two-slit signals is far greater than predicted by a linear summation of individual frequency bands. This pattern holds for three-slit signals, where the XS IQ exceeds 3. In contrast, the XS IQ for manner, voicing and consonant identification is well below 1 for three-slit stimuli. This non-linearity is compressive in nature, meaning that information integration across frequency channels is slightly less than predicted on the basis of linear summation. In view of the fact that consonant recognition improves markedly for three slits (relative to two), this is consistent with place of articulation being the driving phonetic feature underlying consonant recognition. Let us now return to the issue raised earlier in the chapter concerning the contribution made by voicing and manner to intelligibility. Although place of articulation is the most important phonetic feature in ideal listening conditions, voicing and manner are likely to play a crucial role when listening conditions are less than ideal.

WHAT UNDERLIES LINGUISTIC CONTEXT?

Speech communication involves far more than identifying individual segments in isolated syllables and words. Listeners typically decode sequences of words embedded in complex phrasal structures. Linguistic context plays an important role in this process, as Figs. 1 through 3 attest. However, the factors underlying the context effect are still poorly understood. In particular, why is speech so much more intelligible when packaged in phrasal and sentential units? This effect is pronounced in low SNR conditions (Fig. 1).

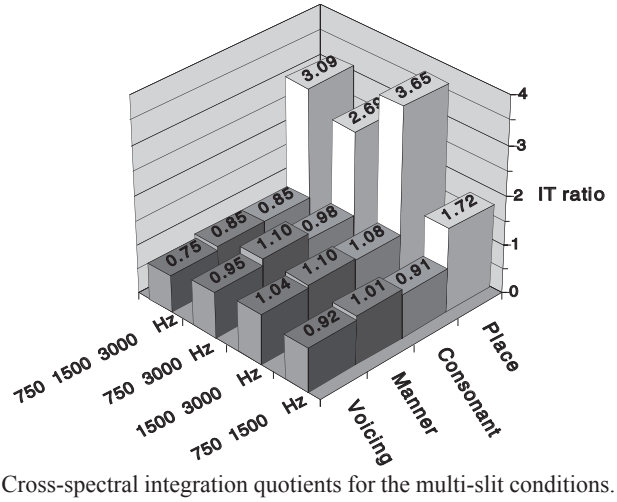

Fig. 7: Cross-spectral integration quotients for the multi-slit conditions. The quotient is defined as the ratio between the observed information transmission for a given multiband condition and the sum of information transmission from the contributing individual bands [adapted from Christiansen and Greenberg (2008)].

This is where prosody may be important. It provides a subsidiary representation that complements the phonetic. Under certain conditions, the phonetic composition of an

utterance may be decodable through prosodic cues that provide sufficient information to deduce the specific identity of segmental elements in an otherwise acoustically compromised signal. It is well known that mental representations of words can be based on prosodic patterns with relatively little phonetic specificity. The tip-of-the-tongue phenomenon (Brown and McNeil, 1966) illustrates this very well. A "missing" word is often mentally tagged by three parameters: (1) the number of syllables, (2) the stress pattern and (3) the initial consonant of the word or primary-accented syllable.

How do the data presented in this chapter relate to prosody? Recall that in Fig. 5, manner and voicing are usually decoded correctly even when consonant recognition is poor. Are voicing and manner of articulation relevant to prosody? We believe this may be the case. This is because voicing and manner are amplitude-contour features that are very sensitive to the flow of energy throughout the syllable. From these features, we believe it is possible for listeners to deduce whether the syllable is stressed or not (or something in between). Moreover, knowing the manner and voicing characteristics of a syllable allows the number of segment alternatives to be pruned significantly. The situation is analogous to a visual image that is partially obscured by an obstruction in the foreground. Often, the object of interest is recognizable despite the visual "noise." Such "glimpsing" allows for the speech signal to be characterized and analyzed under a broad range of listening conditions, many of which are far from ideal.

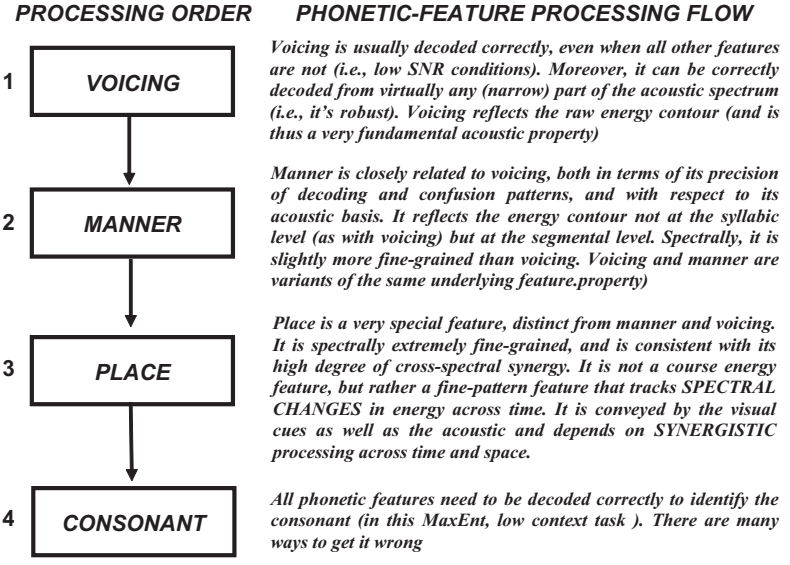

Fig. 8: A schematic illustration of the processing flow of phonetic features from the most coarse to the finest-grain phonetic features.

Another way of examining the consonant recognition data presented in Figs. 5 and 6 is through a conditional probability analysis. When a specific phonetic feature is correctly decoded (and the consonant is recognized incorrectly), what other features are correctly or incorrectly decoded? In our data, when voicing is correctly decoded,

manner is also likely to be accurately decoded, but not so place of articulation. When manner is correctly decoded, voicing is usually decoded as well (but not necessarily place). Manner and voicing decoded are highly correlated with each other, as if they are closely related features.

What happens when a feature is incorrectly decoded? If the feature is place of articulation, it is likely that manner and voicing are correctly decoded. However, if manner is incorrectly decoded, place is unlikely to be decoded accurately. This implies that *decoding place information relies on manner decoding*. Moreover, when voicing is incorrectly decoded, manner is unlikely to be decoded properly, suggesting than *manner is dependent on voicing* analysis. The converse is not the case. An error in manner decoding has relatively little impact on whether voicing is decoded correctly (or not). From such conditional (and highly asymmetric) probability analyses we can delineate the likely flow of processing for consonant recognition. In our view, the processing of phonetic information proceeds from (1) Voicing to (2) Manner to (3) Place of Articulation and finally (4) Consonant recognition (Fig. 8).

Because Voicing and Manner features are (in our view) closely linked to prosody, it is likely that a prosodic analysis is usually performed prior to a detailed phonetic analysis, particularly in uncertain or acoustically challenging listening conditions. We believe that the flow of processing delineated in Fig. 8 is also likely to apply when visual speech cues are present. This is because the visual stream probably requires an energetic (i.e., syllabic) contour of the acoustic signal in order for visual place-of-articulation information to be effectively integrated with manner and voicing cues.

In summary, the process of speech decoding reflects a complex integration of sensory and information streams that often combine in synergistic fashion. The phonetic and prosodic tiers of linguistic analysis are integrated to provide a much richer and more detailed picture than afforded by either representation alone. How these information streams are combined is not well understood. Such knowledge would be extremely useful for understanding how the brain goes from sound to meaning.

REFERENCES

ANSI (**1997**). "Methods for Calculation of the Speech Intelligibility Index," **S3.5-1997.**

Brown, R., and McNeil, D. (**1966**). "The 'tip-of-the-tongue' phenomenon," J. Verb. Learn. Behav., **5**, 325-337.

Chang, S., Wester, M., and Greenberg, S. (**2005**). "An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language," Speech Communication, **47**, 290-311.

Christiansen, T. U., and Greenberg, S. (**2005**). "Frequency-selective filtering of the modulation spectrum and its impact on consonant identification," in The Twenty-First Danavox Symposium, edited by A. Rasmussen and T. Poulsen, 585-599.

Christiansen, T. U., and Greenberg, S. (**2008**). "Cross-spectral synergy is crucial for consonant recognition," Submitted.

Christiansen, T. U., Dau, T., and Greenberg, S. (**2007**). "Spectro-temporal processing of speech – An information-theoretic framework," In Hearing – From Sensory Processing to Perception, edited by B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp and J. Verhey. Berlin: Springer Verlag, 517-523.

French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am., **19**, 90-119.

Grant, K., and Greenberg, S. (**2001**). "Speech intelligibility derived from asynchronous processing of auditory-visual information," Proceedings of the Workshop on Audio-Visual Speech Processing (AVSP-2001), 132-137.

Grant, K. W., and Walden, B. E. (**1996**). "Evaluating the articulation index for auditory-visual consonant recognition," J. Acoust. Soc. Am., **100**, 2415-2424.

Greenberg, S. (**2006**). "A multi-tier theoretical framework for understanding spoken language," In Listening to Speech: An Auditory Perspective, edited by S. Greenberg, and W. A. Ainsworth. Mahwah, NJ: Lawrence Erlbaum Associates, 411-433.

Greenberg, S. (**2007**). "What makes speech stick?" Proceedings of the XVIth International Congress of Phonetic Sciences, 737-740.

Greenberg, S., Arai, T. and Silipo, R. (**1998**). "Speech intelligibility derived from exceedingly sparse spectral information," Proceedings of the 5th International Conference on Spoken Language Processing, 74-77.

Grosjean, F. (**1985**). "The recognition of words after their acoustic offset: Evidence and implications," Percept. Psychophys., 38, 299-310.

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. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am., **27**, 338-352.

Miller, G. A., Heise, G. A., and Lichten, W. (**1951**). "The intelligibility of speech as a function of the context of the test materials," J. Exp. Psych., **41**, 329-335.

Müsch, H., and Buus, S. (**2001**). "Using statistical decision theory to predict speech intelligibility. II. Measurement and prediction of consonant-discrimination performance," J. Acoust. Am. Soc., **109**, 2910-2920.

Pickett, J. M., and Pollack, I. (**1963**). "Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt," Language and Speech, **6**, 151-164.

Plomp, R. (**2002**). The Intelligent Ear. Mahwah, NJ: Lawrence Erlbaum Associates.

Pollack, I., and Pickett, J. M. (1963). "The intelligibility of excerpts from conversation," Language and Speech, **6**, 165-171.

Silipo, R., Greenberg, S., and 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), 2687-2690.