Talker Variability and Normalization

- Abstractionist (Classic) View - Speech perception is a process of abstraction. Variation irrelevant to phonetic representation is “noise” and perception normalizes to account for this variation.
- Abstractionist (Prototype) View - Abstraction matches to an averaged or typical representation.
- Exemplar View - Speech perception preserves token variation as a source of “lawful variation” that is part of the recognition process.

Abstractionist View

Listener recovers abstract (phonetic, gestural) representation intended by talker. Token variation, talker variation, speaking rate, etc. are all “noise” or variability that must be factored out (normalized) in perception.

Across talkers, there is variation in the length (and width) of the vocal tract and in the mass and size of the vocal folds. This introduces variation in the acoustic realization of phonemes (gestures).

Abstractionist View - 2

Perception must normalize this variation and convert the input back into a canonical form (phonemes or gestures).

Perception “throws away” information about the talker to recover the abstract representation intended by the talker.

Normalization

Normalization is designed to map the tokens onto a standardized representation. Indexical properties of the talker (e.g. acoustic qualities that indicate vocal tract length) are used to adjust the acoustic correlates of phonetic quality.

For vowels, there are a number of different schemes, but they generally fall into two groups: intrinsic and extrinsic.

Intrinsic (Vowel) Normalization

Acoustic cues within each token that are correlated with differences across talkers are used as “indexical” cues. They allow acoustic correlates such as the formant frequencies to be mapped onto a standardized set of coordinates or representation (see Syrdal & Gopal, 1986).

For example, since F0 is correlated with vocal tract length, using F1-F0 in place of F1 as a cue to vowel height reduces the variability due to talker.

Intrinsic - Part 2

Since the talkers vocal tract length is present in F1, F2, and F3, then using F2-F1 and F3-F2 in place of F2 and F3 subtracts out the (constant) contribution of the talker and leaves talker independent cues to vowel identity.

F4 is also a talker dependent (but not phonetic dependent) “cue” that can be used to normalize for the length of the talker’s vocal tract.

Evidence

Create a series of synthetic vowels varying in the values for F1, F2, and F3 (e.g. see Johnson, 1990). The values range from those of /ʊ/ to /ʌ/. Make one series with an F0 of 120 Hz and the other with an F0 of 240 Hz.

When presented in a blocked design (only F0=120 Hz in one block, F0=240 Hz in the other), there was a large difference in labeling. With 240 Hz F0, most stimuli labeled / ʊ / while with a low F0, most labeled /ʌ/.

This corresponds to what would be expected based on a standardized model and male and female talkers. Since listeners used information internal to the token, this is consistent with intrinsic normalization.

Extrinsic (Vowel) Normalization

Listeners accumulate information across tokens about the range of F1, F2, and F3. This information establishes the talker’s vowel space which can then be mapped onto a standard. The mapping rule can then be applied to all new tokens to map them to a standard form.

For example, if the range of F1 is established as 300 to 700 Hz, a token with an F1 at 500 is at .5 of the F1 space (half-way between minimum and maximum). Gerstman (1968) proposed this type of scheme for vowel normalization.

Since each vowel is specified in terms of a standardized (like a z-score) space, talker differences are eliminated.

Evidence - 2

Create ambiguous tokens (e.g. between /ɪ/ and /ɛ/). Create synthetic sentences, with F1 in different F1 ranges. Sentence is followed by a token (/bɪt/ or /bɛt/) for a listener to identify (Ladefoged and Broadbent, 1957).

Following the low F1 sentence, ambiguous tokens are classified as “lower” vowels (high F1, more /ɛ/ responses). Following the higher F1 sentence, more “high” (low F1, /ɪ/) responses. Also found effects of F2.

This is consistent with extrinsic normalization since listeners are clearly using the information in the sentence in their recognition of the vowel (word) target.

Exemplar View

Listeners extract the voice information and the token information simultaneously, from overlapping acoustic cues.

The acoustic cues are mapped to an exemplar based memory representation that preserves detail (indexical information). This detail is part of how the system recognizes phonemes and words.

When one is listening for meaning, the indexical properties are not the focus of attention, but they are still part of the recognition process.

Exemplar View - 2

Within this approach, we expect that:

1) Classification is exemplar rather than prototype based.
2) Phonemes, words and sentences in a familiar voices should be easier to recognize than an unfamiliar voice.
3) Memory tasks will show retention of detail and other phenomena related to exemplar memory.

Evidence - 3

Data from Pisoni, Nygard and collaborators shows:

1) After training on talker recognition, novel words in the same voice are better recognized than the same novel words in a novel voice. (Same for sentences.)

2) When listeners are given a task (e.g. syllable or word recognition) with multiple talkers, their performance is worse than with a single talker.

3) The perceptual processing of phonetic quality and talker voice are integral.

Interpretation

The data for intrinsic or extrinsic normalization can be captured within an exemplar model.

The data on integrality can be used to argue that the same information underlies both talker identity and phoneme identity and that they are not separate, independent representations.

The data on voice familiarity are consistent with exemplars, but also some normalization models.

The data on talker variability do not distinguish models.

Some Further Alternatives

The system may compute both an abstract representation and preserve detail. The time course for the processing of details may be different than the representation needed for primary word recognition (McLennan).

In work on memory, there is some evidence for separate streams of information that have dissociable properties. Explicit memory tasks seem to tap the abstract representation. Implicit memory tasks show evidence that details (exemplars?) have been preserved.