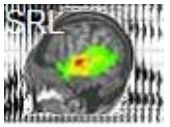
Integration of Speech- and Voice-Perception Systems for Native Language Talker Identification

Tyler K. Perrachione[1,5] Joan Y. Chiao[2,5,7] Todd B. Parrish[3,6] Janet B. Pierrehumbert[1,5,8] Patrick C.M. Wong[4,7,9]

[1]Department of Linguistics, [2]Department of Psychology, [3]Department of Biomedical Engineering, [4]Roxelyn & Richard Pepper Department of Communication Sciences and Disorders, [5]Cognitive Science Program, [6]Department of Radiology, [7]Northwestern University Interdepartmental Neuroscience Program, [8]Northwestern Institute on Complex Systems, [9]Program in Computational Biology and Bioinformatics, Northwestern University, Evanston, Illinois, U.S.A

pwong@northwestern.edu



Speech Research Laboratory
http://www.soc.northwestern.edu/wong/

Abstract

This study demonstrates increased functional integration between brain regions responsible for speech and voice perception during the identification of talkers speaking a native versus foreign language. Behavioral studies have demonstrated that individuals are more accurate at identifying voices when they understand the language being spoken. Previous neuroimaging studies of voice perception have predominantly contrasted either speech vs. non-speech or voice vs. verbal content. However, these contrasts preclude detecting ways in which speech- and voice-perception systems work in conjunction in a talker identification task. By contrasting voice identification in a native vs. foreign language, we demonstrate that neural systems responsible for encoding auditory information onto meaningful structure (i.e. the phonology of one's native language) contribute more to identification of voices in one's native language.

Introduction

Understanding speech requires mapping between high-variability auditory information (intra- and inter-speaker variation) and meaningful (invariant) phonological units.

Variability due to voice has a well-established effect on speech perception abilities.

Specific linguistic experience affects speech perception abilities.

Listeners demonstrate a "Language-Familiarity Effect" on talker identification tasks. Voices speaking a familiar language are more accurately identified than those speaking an unfamiliar language. (Goggin et al., 1991)

Perrachione & Wong (2007a): The Language-Familiarity Effect is a True Linguistic Effect:

> **Exposure to foreign-language voices not enough; specific linguistic knowledge (proficiency) is necessary to over come the Language-Familiarity Effect.**
>
> **Listeners of different language backgrounds find different voices more confusable – may attend to different cues.**
>
> **Suggests a bidirectional integration between speech & voice.**

Neurologic evidence from talker identification in a dichotic listening paradigm provides early evidence confirming this link. Right-ear (left cerebral hemisphere) accuracy is a better predictor of overall accuracy when identifying voices in a native versus foreign language (Perrachione & Wong, 2007b)

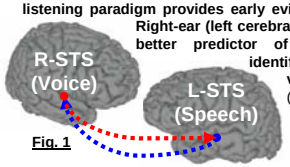


Fig. 1

The right superior temporal sulcus (STS) is the primary locus of the human voice perception system (Belin et al., 2000).

However, methodological issues have prevented prior neuroimaging studies from being able to demonstrate the functional connection between speech and voice perception systems:

Stimulus-Based Approaches contrast activation due to speech vs. non-speech (Belin et al., 2000; Fecteau et al., 2004).

> **This approach cannot reveal a functional integration because speech and voice are confounded into a single condition.**

Task-Based Approaches contrast activation due to attending vocal identity vs. verbal content (Stevens, 2004; von Kriegstein et al., 2003; von Kriegstein & Giraud, 2004).

> **This approach cannot reveal a functional integration because identical speech information is present in both conditions.**

Methods

Subjects

- **8 native speakers of American English**
- **No prior knowledge of Mandarin Chinese**
- **6 females / 2 males, age 18-24 years (M = 21.0)**
- **All right-handed, and free from hearing or neurologic deficits**

Stimuli

- **2 Language Conditions (English, Mandarin)**
- **3 Voices / Language Condition**
- **4 Training Sentences; 12 Test Sentences**
  - *e.g.* "A rod is used to catch pink salmon"
  - *e.g.* "这是一个美丽而神奇的景象"
- **Normalized to 70dB SPL RMS Intensity**
- **Matched for duration**
  - $M_{English}$ = 2.330s, $M_{Mandarin}$ = 2.398s [$t(70) = -1.208$, $p = 0.231$]

Procedures

1. **Trained to recognize 3 voices in scanner with feedback (~7:30min, 72 trials)**
2. **Tested on ability to recognize voices**
   - **2 Test Runs / Language Condition**
   - **36 Trials, Jittered Presentation Onset**
   - **254 TRs / Run**
3. **Repeat in other Language Condition**

Scanning Parameters

Functional Volumes	Anatomical Volume
- T2*-weighted EPI pulse	- Axial orientation
- TR/TE = 2s / 30ms Flip angle = 90º	- T1-weighted MP-RAGE
- Slice Thickness 3 × 3.4 × 3.4mm	- TR/TE = 2.1s / 2.4ms
- 30 slices / Volume	- Flip angle 15°

Preprocessing

Discard first 5 TRs, motion correction, align to first functional volume, spatial smoothing (6mm FWHM), normalized for percent signal change, detrended, and resampled to 3mm³ voxels.

MRI preprocessing and data analysis conducted with AFNI.

Results

Data were analyzed in a mixed-effects 3D ANOVA, with Condition as a fixed factor and Subject as a random factor

All Trials:

Contrasting activation from attending English Voices > Mandarin Voices revealed a significant cluster in Left mSTS ("Phoneme Area") (e.g. Liebenthal et al., 2005)

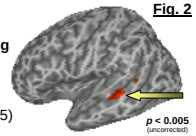
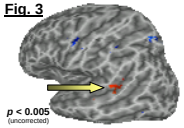


Fig. 2

Fig. 3



Correct Trials:

English Voices > Mandarin Voices also revealed greater activation in the English condition in left mSTS when to accurate response trials.

English Trials:

Correct Trials > Incorrect Trials reveals greater activation in the anterior STG temporal pole) in response to correct responses

This area appears frequently in voice perception studies (e.g. Imaizumi et al., 1997)

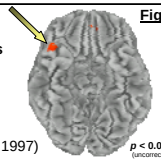


Fig. 4

Region of Interest (ROI) Analyses

Percent signal change (unthresholded) was extracted from all subjects in two functional ROIs (5 voxel radii) centered on the most activated voxel in Fig. 2 (Left-mSTS) and to coordinates from Belin et al. (2000) for the human voice selective area (Right-mSTS).

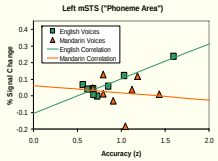


Fig. 5: Activity in Left mSTS was a significantly better predictor of accuracy when subjects listened to English voices than Mandarin voices.

$r_{English}$ = 0.912 $r_{Mandarin}$ = -0.113

[$z = 2.612$, $p < 0.0045$]

Fig. 6: The predictive capacity of Right mSTS activity (based on Belin et al. 2000's "Voice Area" coordinates) did not differ between conditions.

$r_{English}$ = 0.900 $r_{Mandarin}$ = 0.543

[$z = 1.370$, *n.s.*]

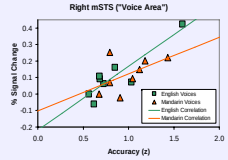
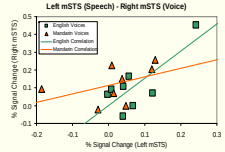


Fig. 7: The correlation between activity in Left mSTS and Right mSTS was significantly stronger in the English versus Mandarin condition, suggesting a stronger functional integration between these regions during native-language talker identification

$r_{English}$ = 0.784 $r_{Mandarin}$ = 0.451

[$z = 1.669$, $p < 0.048$]



Discussion

This study demonstrates a functional integration between left-hemisphere language (speech perception) areas (Left mSTS) and right-hemisphere voice perception areas (Right mSTS) for a talker identification task.

Left mSTS is significantly more activated by native- versus foreign language talker identification.

Activity in Left mSTS is more strongly correlated with accuracy identifying talkers in a native language, whereas Right mSTS activity does not differ in its predictive capacity of native- versus foreign-language talker identification accuracy.

Activity in Left mSTS and Right mSTS are significantly more correlated in a native versus foreign language, suggesting greater functional integration of these regions for language talker identification.

Selected References

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P. & Pike, B. (2000) "Voice-selective areas in human auditory cortex." *Nature*, 403, 309-312.

Fecteau, S., Armony, J.L., Joanette, Y. & Belin, P. (2004) "Is voice processing species-specific in human auditory cortex? An fMRI study." *Neuroimage*, 23, 840-848.

Goggin, J.P., Thompson, C.P., Strube, G. & Simental, L.R. (1991) "The role of language familiarity in voice identification." *Memory and Cognition*, 19, 448-458.

Imaizumi, S., Mori, K., Kiritani, S., et al. (1997) "Vocal identification of speaker and emotion activates different brain regions." *NeuroReport*, 8, 2809-2812.

Liebenthal, E., Binder, J.R., Spitzer, S.M., Possing, E.T. & Medler, D.A. "Neural substrates of phonemic perception." *Cerebral Cortex*, 15, 1621-1631.

Perrachione, T.K. & Wong, P.C.M. (2007a) "Learning to recognize speakers of a non-native language: Implications for the functional organization of auditory cortex." *Neuropsychologia*, 45, 1899-1910.

Perrachione, T.K. & Wong, P.C.M. (2007b) "Increased left-hemisphere contribution to native-versus foreign-language talker identification revealed by dichotic listening." *16th Meeting of the International Congress of Phonetic Sciences* (Saarbücken, Germany, August 2007).

Stevens, A.A. (2004) "Dissociating the cortical basis of memory for voices, words and tones." *Cognitive Brain Research*, 18, 162-171.

von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A.L. (2003) "Modulation of neural responses to speech by directing attention to voices or verbal content." *Cognitive Brain Research*, 17, 48-55.

von Kriegstein, K. & Giraud, A.-L. (2004) "Distinct functional substrates along the right superior temporal sulcus for the processing of voices." *Neuroimage*, 22, 948-955.

The authors extend warm thanks to Geshri Gunasekara, Ajith Kumar Uppunda, Nondas Leloudas, Tasha Dees, Louisa Ha, Anil Roy, James Jin, and Neha Malhorta for their assistance in this research. This work is supported by Northwestern University and the National Institutes of Health (U.S.A.) grants HD051827 & DC007468 to PW.