A phonetically-based phonetic similarity metric

Jeff Mielke

October 31, 2009

1 Introduction

This talk demonstrates a new metric for measuring phonetic similarity based on several types of phonetic data collected for a wide range of consonants and vowels. Phonetic similarity is frequently invoked for explanating a wide range of phonological observations, and the purpose of this project is to provide a resource for quantifying phonetic similarity, distinguishing different types of phonetic similarity (articulatory, acoustic, and perceptual) and for distinguishing phonetic similarity from phonological notions of similarity, such as those based on features (e.g. Frisch (1996); Frisch et al. (2004), Kondrak (2003, *et seq.*)) or on phonological patterning. This paper describes the measured phonetic similarity between a set of crosslinguistically frequent consonants and vowels, and compares this with a coarse measure of phonological similarity that is based on the crosslinguistic patterning of these sounds.

The original motivation for creating the similarity metric was to investigate the role of phonetic similarity in determining the sets of segments that are involved in sound patterns. It has long been known that the phonologically active classes appearing in languages' sound systems are not randomly assembled, but reflect general crosslinguistic tendencies. Distinctive feature theory, at its peak (Chomsky and Halle, 1968; Clements, 1985; Clements and Hume, 1995, *inter alia*), was a model of the phonetic parameters used to form contrasts and define natural classes. More recently, it has been argued that an all-purpose model is in conflict with many of the attested phonologically active classes, including recurrent ones (Mielke, 2008a), and that the preference for certain classes (e.g., vowels, nasals) is driven by physiological and perceptual factors and their role in diachronic change (Ohala, 1981; Blevins, 2004, *inter alia*). The goal of this project is to investigate the role of ubiquitous phonetic factors in accounting for which sounds tend to pattern together and which sounds do not.

2 Phonetic similarity

2.1 Methods

2.1.1 Scope of project

Data come from four trained phoneticians (native speakers of English, one female and three males) who each produced the same 58 crosslinguistically-frequent consonants and vowels. An additional set of 72 less frequent sounds were divided up among the speakers and produced by one speaker each. These sounds are shown in Figure 1. Each speaker produced a total of nine repetitions of each sound in three contexts (3 × a__a, 3 × i__i, 3 × u__u). The analysis in this paper will focus on a subset of these sounds.

ʘ | ‖ ! ǂ
p pʲ pʷ   **t** tʲ   ʈ c **k** kʲ **kʷ** k͡p   q qʷ   **ʔ**
p' pʰ   t' tʰ   k' kʰ kʷ'   q'
b bʲ bʷ   **d** dʲ   ɖ ɟ **g** gʲ gʷ g͡b
ɓ ᵐb   ɗ ⁿd   ᵑg
                                 ts   **tʃ**
                                 ts' tsʰ   tʃ' tʃʰ
                                 dz   **dʒ**
ɸ   **f** fʲ θ **s** sʲ ɬ   **ʃ** **x** xʲ xʷ   χ χʷ ħ **h** hʲ hʷ
β   **v** vʲ ð **z**   zʷ   **ʒ** ɣ   ʁ   ʕ
m mʲ mʷ   **n** nʲ   ɳ **ɲ** **ŋ**   ŋʷ
                                 l lʲ   ɭ ʎ
                                 ɾ ʈ
        ʋ   **r** rʲ   ɻ **j**   **w**

                 i **iː** ĩ y ɨ ɨː   ɯ **u** **uː** ũ
                 ɪ   ʊ
                 e **eː** ẽ ø ʌ   **o** **oː** õ
                 ɛ **ɛː** ɛ̃ œ ə   **ɔ** **ɔː** ɔ̃
                 æ   **a** **aː** ã ɑ

Figure 1: Consonants and vowels recorded for this project (**bold** segments produced by all speakers)

2.1.2 Data collection

Each speaker was recorded in two sessions, because some of the data collection techniques are incompatible with each other. Session 1 involved audio, video, and ultrasound tongue imaging, and session 2 involved audio, electroglottography, and airflow measurement.

In session 1, audio was recorded using an Audio-Technica PRO 49Q condenser microphone and a Symetrix 302 microphone preamplifier. The tongue was imaged using a SonoSite TITAN portable ultrasound machine with a C-11/7-4 11-mm broadband curved array transducer placed under the chin, generating a mid-sagittal section from near the tongue root to near the tongue tip at a rate of 28 scans per second, output as 29.97 fps analog video.

The palate was imaged at the beginning of the session by asking the speaker to drink water through a straw. The speaker's face was recorded in profile with a Sony Mini-DV Digital Handycam, and this video was used to record lip movements and to generate data for movement correction (described below). The audio channel and the two video channels (ultrasound and camcorder) were combined and digitized using a Videonics MXProDV digital video mixer. The audio was digitized at a sampling rate of 48Hz with 16 bits per sample. The result of combining the video channels is shown in Figure 2. The tongue surface is the bright white contour in the fan-shaped ultrasound view (the sound is [a]), and the four pink dots are on blue sticks attached to the glasses and the ultrasound transducer. A blue screen behind the speaker allows everything but the dots and the speaker's face to be removed from the camcorder video.
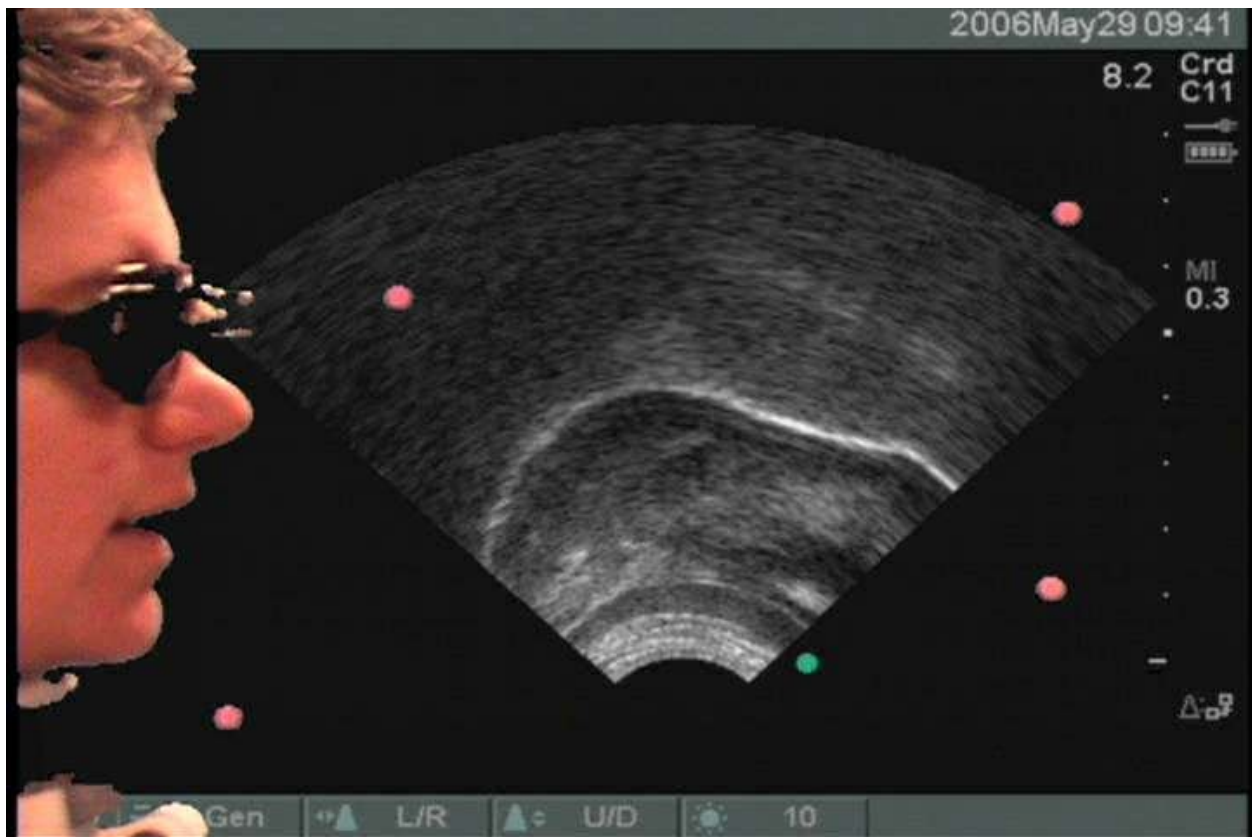
Figure 2: Video data collection

In session 2, oral and nasal airflow were measured using a SCICON R&D Macquirer 516 transducer interface system and the associated oral and nasal masks. Vocal fold contact area (a measure of voicing) and larynx height were measured using a Glottal Enterprises EG2-PC electroglottograph connected to two data channels of the Macquirer 516. The four channels of articulatory data were recorded at a sampling rate of 1100Hz using the Macquirer software. Audio was recorded using the microphone built into the oral mask but used only as an index for the articulatory channels, and not for further analysis of the audio itself. The Crackquirer program (Baker, 2006) was used to convert the data from a proprietary format

into wav files and tab-delimited text files for the articulatory data.

2.1.3   Data analysis

All of the recordings from both sessions were segmented using TextGrids in Praat (Boersma and Weenink, 2007), which were used for all subsequent measurements and comparisons in all of the channels.

Acoustic distances between segments were calculated using a dynamic time warping algorithm (Holmes and Holmes, 2001) implemented in Python. Waveforms for each token (each VXV sequence produced by the speakers) were first converted into matrices of 12 Mel-frequency cepstral coefficients and 12 delta coefficients based on 15ms windows at 5ms intervals, using Praat. For each pair of segments, the distances were normalized by speaker and then averaged across speakers, contexts, and repetitions, to produce one acoustic distance for each pair of segments.

The articulatory data from session 2 (airflow and EGG) was measured as follows. For larynx height and for oral and nasal airflow, the data were first smoothed by averaging across 3ms windows, and then average value for each target segment interval was divided by the average value for the preceding context vowel interval. For vocal fold contact area, the absolute value of the extremum during the target interval was divided by the absolute value of the extremum during the preceding vowel interval. The measured values (in arbitrary units) are plotted in Figures 4 and 5[1], below.

To generate measurements of the shape of the vocal tract, video frames were extracted during the target segment intervals and during palate imaging. Because of the video (NTSC) is 29.97 frames per second, frames are about 33 ms apart. Palatoglossatron (Baker, 2005) was used to semi-automatically trace the tongue contour and lips in each tongue image, and the palate contour in each palate image. In all, about 13,000 video frames were traced by a research team. The Palatron algorithm (Mielke et al., 2005) was used to transform all of the traces in order to compensate for head/transducer movement, so that the tongue, palate, and lips could be combined in the same coordinate system. A Python script was used to interpolate between the lips and the anterior extent of the tongue and palate images, to estimate the location of the pharyngeal wall, and to generate cross-distances between the upper and lower vocal tract surfaces at 5mm intervals. The measured version of the image in Figure 2 is shown in Figure 3. This process generates a vector for each image, representing the measured/estimated shape of the vocal tract.

The data were normalized across three ways, to make them comparable. The vocal tract was normalized lengthwise by dividing the vocal tract into zones (pharnyx, upper pharynx/velum, hard palate, alveolar ridge, and lips) and resampling the cross-distances so that every image has the same number of measurements per zone, resulting in 63 measurements for the length of each vocal tract. This has the side-effect of eliminating differences that change the length of the vocal tract, such as lip protrusion. The cross-distances were normalized across speakers by measuring the average distance between the upper and lower surfaces and adjusting scaling all the measurements for each speaker so that the averages are the same for each speaker. Since the target segments differ in duration, the cross-distances

---

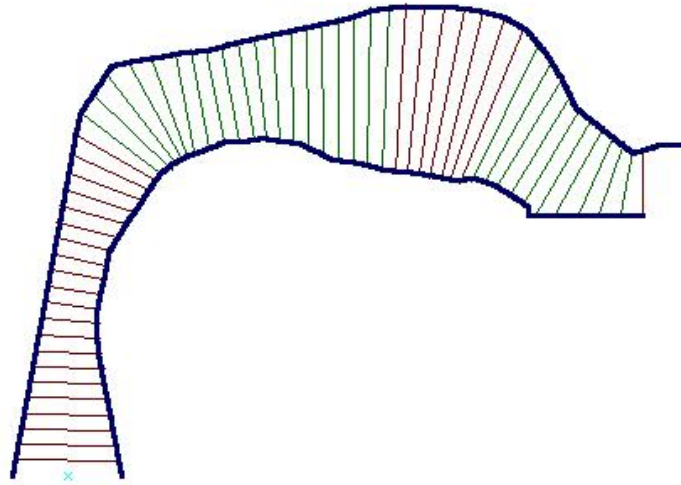[1]/r/ is not shown in the figure. Its VFCA value of over 4 is probably anomalous

Figure 3: Cross-distances automatically measured for measured/estimaged vocal tract

were resampled in the time dimension so that each token is represented by five frames. The result of these steps is that each token is represented by a 5×61 matrix, which can be treated as a 305-dimensional vector.

Zharkova and Hewlett (2008) use ultrasound to show that the effect of neighboring vowels on the tongue shape of consonants is greater than the effect of neighboring consonants on the tongue shape of vowels. The present ultrasound data are no exception to this. Performing a regular principal component analysis on these tokens causes different consonants produced in similar contexts to cluster together more noticeably than similar consonants produced in different contexts. Algorithmically-reweighted principal component analysis (Mielke and Roy, 2009) was used to give more weight to parts of the vocal tract that are important for particular target sounds. Vocal tract positions with relatively high variance across contexts are set to the overall average for all segments. Consequently, only relatively invariant vocal tract positions contribute to the differences between different segments. For example, the vocal tract shapes for the [p]s in [apa], [ipi], and [upu] have nothing in common except for the lips. Rather than identify three distinct tongue shapes as important for articulating [p]s, algorithmic reweighting essentially ignores the non-lip portions of the vocal tract on the basis of their variability across contects. A principal component analysis was performed on the reweighted vectors, after tokens of each segment were averaged.
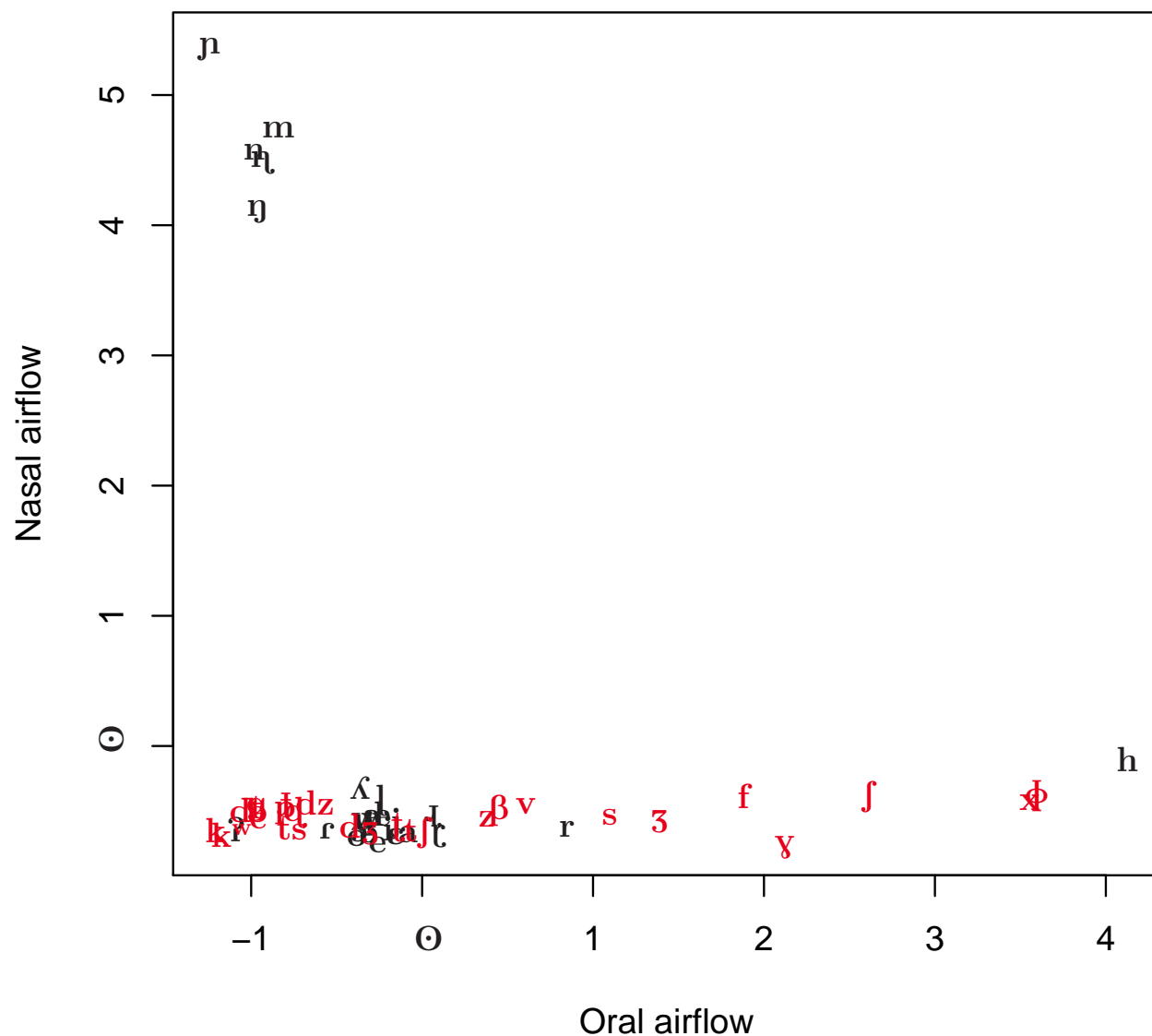
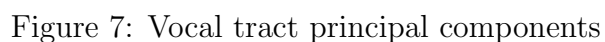### 2.1.4   Data not appearing in this paper

An additional perceptual similarity component of the project has not yet been completed, and all of the data from one male speaker had to be excluded from the analyses reported here due to a fixable problem related to ultrasound data extraction, and the data from long vowels, [ɯ], and the segments produced by only one speaker have been withheld for the time being due to issues with integrating them into the distance metric. In the future, ultrasound

frames from the context vowel intevals will be traced and included in the analysis. The intent is for the similarity metric to be useful now, and to increase in its scope and detail as data analysis proceeds.

2.2 Results

The phonetic distances between these sounds, as measured by these techniques, are displayed in Figures 4-6. The figures were generated by a Similarity Calculator application which generates plots of user-selected segments, and identifies dimensions in which a user-selected subset of these is distinct from other segments. For illustration, the obstruents, as traditionally defined (Chomsky and Halle, 1968), are colored red.

Figure 4: Airflow measurements

Figure 5: Electroglottography measurements

The first two figures show the directly-measured articulatory dimensions. Figure 4 shows, not surprisingly, that nasal airflow distinguishes nasals from the other sounds. Oral airflow distinguishes fricatives (including [h]), and also [r], from other sounds. Figure 5 shows that voiceless obstruents, voiced obstruents, and sonorants are generally distinguished by vocal fold contact area. Voiced stops are distinguished from sonorants and other obstruents by larynx height. These phonetic measures group glottals [ʔ h] with obstruents instead of

Figure 6: Acoustic principal components

sonorants. Voiced fricatives, especially [ɣ], show an affinity to sonorants.

The next two figures show the segments plotted according to the first two principal components of the acoustic and vocal tract data. Figure 6 shows that the eight sibilants (which share high-frequency noise) are distinguished from the other segments by the first principal component, which seems to correlate with spectral tilt or stridency vs. sonority. The non-sibilant obstruents are distinguished from the sibilants and most of the sonorants along the second principal component, which seems to correlate with amplitude. The glottals [ʔ h] again group with obstruents, as do the flaps and [ɨ]. Differences between vowels and sonorant consonants, which here seem scrambled, are apparent in the higher principal components (not pictured). Differences in place of articulation generally are not.

Place of articulation is more apparent in the vocal tract shape data in Figure 7. Here the

Figure 7: Vocal tract principal components

first principal component is correlated with the opposition between palatal and pharyngeal constrictions. Palatal consonants and vowels are at one extreme, opposite the pharyngeal and lax vowels, along with [r ɻ]. The second principal component is correlated with velar vs. more anterior lingual constrictions, distinguishing the velar stops and [u] from the other segments. The pile of segments close to the point (0,0) in the figure contains labials, glottals, and [ɾ]. Because the first two principal components involve constriction locations which the reweighting algorithm does not deem important for any of these sounds, their cross-distances have been set to near the average for all segments. The labials are distinguished by higher principal components which involve the lips (not pictured), but the glottals are not distingished from any of the other sounds by supralaryngeal vocal tract shape. The odd behavior of [ɾ], as well as [r ɻ], may be explained by methodological issues, discussed below

in the discussion.

3   Phonological similarity

To compare the phonetic distances with phonological patterning, a measure of phonological similarity was formulated using the sound patterns reported in P-base (Mielke, 2008a). The database contains 6077 phonologically active classes (classes that serve as the trigger or target for an alternation). Many of these classes are involved in multiple sound patterns within the same language. The availability of these data in electronic form provide an easy way to get a coarse measure of similar phonological patterning.

3.1   Methods

Phonological similarity was measured by counting the cooccurrences of pairs of sounds in the same phonologically active classes. Specifically, the phonological similarity between two sounds equals the number of sound patterns involving both segments in the same class, divided by the number of sound patterns involving either segment (in languages which have both sounds in the inventory). The value would be 1 for two sounds that always pattern together, or 0 for two sounds that never pattern together but do occur in the inventories of the same languages. A principal component analysis of the phonological distances ($1 - similarity$) was performed.

3.2   Results

The first principal component distinguishes vowels from consonants. The vowels are very close together due to many sound patterns involving the class of vowels. Since they are clearly separated from the other segments, it makes sense to analyze them separately to look for structure internal to this cluster, but that is not pursued here. The second dimension generally distinguishes sonorant consonants from obstruents. The middle ground is occupied by the glottals, the nonsibilant voiced fricatives (except [β], which is clearly situated among the sonorants), and [ɸ].

4   Discussion of results

Some familiar patterns are seen in the phonetic and phonological similarity reported above, involving the patterning of nonsibilant voiced fricatives, glottals, and the association of particular contrast with particular types of data. Some strange behavior may be attributed to methodological issues.

Nonsibilant voiced fricatives often appear in the figures with sonorant consonants or between obstruents and sonorant consonants. This is especially noticeable in the EGG data (Figure 5), where voiced fricatives are ambiguous and [ɣ] is among the sonorants, and the phonological similarity plot (Figure 8), where voiced nonsibilant fricatives and [ɸ] are ambiguous and [β] is among the sonorants. The fact that this pattern is seen in both

Figure 8: Phonological similarity (phonologically active classes)

phonetic and phonological data is consistent with the idea the ambiguous phonetics leads to ambivalent phonological patterning across languages (Mielke, 2005a).
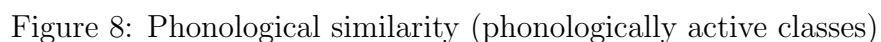
Glottals [ʔ h] are another group of historically ambiguous segments. The SPE definition that groups them with sonorants defines the opposition in terms of the compatibility of the "vocal tract cavity" with spontaneous voicing (Chomsky and Halle, 1968, 302). The only phonetic measure that directly relates to this is the vocal tract shape measure (Fig 7), where they are located with other sounds that do not require a lingual constriction. Acoustic similarity 6 and EGG 5 put them closer to obstruents, and phonological patterning (Figure 8), which presumably is sensitive to a wide range of phonetic dimensions, puts them in the middle. The intermediate location of the glottals in the phonological similarity figure is also consistent with Ratelle (2009)'s investigation of specific cases where glottals pattern with

traditionally [+sonorant] or [−sonorant] sounds.

The association of acoustic similarity with major class and manner distinctions and articulatory similarity with place distinctions is consistent with Lin and Mielke (2007)'s argument that manner distinctions may be learned more easily from acoustic data while place distinctions may be learned more easily from articulatory data.

The strange patterning of the trill and flaps [r ɾ ɽ] in the vocal tract data are likely due to the fact they involve closures shorter than the 33ms interval between the video frames. They are difficult to observe without high-speed ultrasound techniques (Miller and Finch, 2009). Another factor may be that these sounds do not correspond to native phonemes for the speakers who produced them, and this could account for [r]'s strange behavior in non-vocal tract dimensions as well. The trill requires a precise aerodynamic configuration that most of the other sounds do not require. It is possible that the speakers (who as linguists have been trained to produce non-native sounds) are able to produce sounds that are acceptable auditorily, but are distinguishable from natively-produced sounds by the methods used here. One of the reasons for collecting different sets of sounds from different speakers in this project was to prepare for a future project in which different sets of sounds are produced by native speakers of languages containing those sounds.

5   General discussion

A basic question in phonological theory is **Why is the class of obstruents active in many different languages?** Many similarly-phrased questions can be asked about other groups of sounds. The leading answer to this question may once have been **Because the feature [sonorant] is in Universal Grammar** . In the absence of this UG-based explanation, a comparable answer would be **Because obstruents are phonetically similar to each other**.

Supporting this answer requires a means of quantifying phonetic similarity, which is the objective of the project described here. This paper has illustrated a few different ways in which obstruents are phonetically and phonologically distinct from other sounds. It has also been seen that the phonetic and phonological distinctions are gradient in similar ways. Thus there is a phonetic basis for the phonological observations.

While there is a clear connection between phonetic similarity and phonological activity, it is worthwhile to reflect on the relationship. A more sophisticated answer to the question would provide a specific mechanism for how a phonetic fact becomes incorporated into a phonological grammar. Data about the phonological patterns that involve the phonetic distinction provide some insight: [−sonorant] is needed about three times as often [+sonorant] in P-base sound patterns (Mielke, 2008b), and of 81 instances of the class [−sonorant] in the database, about half involve voicing or devoicing patterns. An important mechanism for the involvement of the sonorant-obstruent opposition in phonology seems to be the tendency of adjacent obstruents to affect each other's voicing specification, due to the difficulty of producing adjacent voiceless and non-spontaneously voiced intervals. In this case phonetic similarity is only indirectly related to the mechanism of change.

In light of this, a more direct way of accounting for recurrent classes would involve a comparing the record of sound changes with the record of synchronic sound patterns,

something which will be possible upon completion of the Handbook of Phonological Change (Blevins, in prep), a database of regular sound changes.

A more direct role for phonetic similarity may be generalization in the acquisition or spread of sound patterns (Mielke, 2005b, 2009; Dinkin, 2006, *inter alia*). Mielke (2005b) argued that a tendency to associate phonological behavior with phonetically similar segments results in a bias toward phonetically natural classes. This could account for sound patterns involving classes that do not appear to be related to the original phonetic basis for the pattern. The extent to which generalization is needed to account for natural classes depends on the magnitude of the residue after sound patterns with direct phonetic explanations have been removed.

If the direct role of phonetic similarity is limited to generalization, then it is reasonable to suppose that it would involve phonetic dimensions that are salient to the language user. The salience of articulatory similarity, especially the vocal tract measure in this project, is suspect. Future studies may show that other phonetic dimensions are dominant in generalization, or that the role of articulatory similarity is better described in terms of tactile feedback or muscular activity.

6 Conclusions

This paper has described a phonetic similarity metric that is based on phonetic data, and compared its results with a measure of phonological similarity. Phonetic similarity is an important factor to consider in accounting for sound patterns, and ideally would be entertained prior to positing more elaborate notions of similarity, but quantifying phonetic similarity is a prerequisite for this. Phonetic similarity needs to be considered in terms of specific mechanisms by which phonetic factors can come to affect a phonological system.

References

Baker, Adam. 2005. *Palatoglossatron 1.0*. University of Arizona, Tucson, Arizona. URL http://dingo.sbs.arizona.edu/~apilab/pdfs/pgman.pdf.

Baker, Adam. 2006. *Crackquirer (software)*. University of Arizona, Tucson, Arizona. URL http://dingo.sbs.arizona.edu/~apilab/pmwiki/pmwiki.php?n=Procedures.ConvertingFlowAndPressureDataToPraatFormat.

Blevins, Juliette. 2004. *Evolutionary Phonology*. Cambridge: Cambridge University Press.

Blevins, Juliette. in prep. The handbook of phonological change.

Boersma, Paul, and David Weenink. 2007. *Praat: doing phonetics by computer [computer program]*. URL http://www.praat.org.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English*. Cambridge, Mass.: MIT Press.

Clements, G. N. 1985. The geometry of phonological features. *Phonology Yearbook* 2:225–252.

Clements, G.N., and Elizabeth V. Hume. 1995. The internal organization of speech sounds. 245–306. Cambridge Mass.: Blackwell.

Dinkin, Aaron. 2006. Unnatural classes and phonological generalization in dialect formation. Talk presented at NWAV 35, Columbus, Ohio.

Frisch, Stefan. 1996. Similarity and frequency in phonology. Doctoral Dissertation, Northwestern University.

Frisch, Stefan, Janet Pierrehumbert, and Michael Broe. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22.

Holmes, John, and Wendy Holmes. 2001. *Speech synthesis and recognition, 2nd edition*. New York: Taylor & Francis.

Kondrak, Grzegorz. 2003. Phonetic alignment and similarity. *Computers and the Humanities* 37:273–291.

Lin, Ying, and Jeff Mielke. 2007. Discovering place and manner features: What can be learned from acoustic and articulatory data? In *Proceedings of 31st Penn Linguistics Colloquium*.

Mielke, Jeff. 2005a. Ambivalence and ambiguity in laterals and nasals. *Phonology* 22.2:169–203.

Mielke, Jeff. 2005b. Modeling distinctive feature emergence. In *WCCFL 24*.

Mielke, Jeff. 2008a. *The Emergence of Distinctive Features*. Oxford: Oxford University Press.

Mielke, Jeff. 2008b. Phonologization and the typology of feature behavior. Paper presented at the Phonologization Symposium, University of Chicago.

Mielke, Jeff. 2009. Accepting unlawful variation and unnatural classes: a model of phonological generalization. In *Variation and gradience in phonetics and phonology*, ed. Ruben van de Vijver Caroline Féry and Frank Kügler. Berlin: Mouton de Gruyter.

Mielke, Jeff, Adam Baker, Diana Archangeli, and Sumayya Racy. 2005. Palatron: a technique for aligning ultrasound images of the tongue and palate. In *Coyote papers vol. 14*, ed. Scott Jackson and Daniel Siddiqi.

Mielke, Jeff, and Joseph Roy. 2009. Measuring articulatory similarity with algorithmically reweighted principal component analysis. Poster presented at the 157th meething of the Acoustical Society of America, Portland.

Miller, Amanda, and Ken Finch. 2009. Corrected high-speed anchored ultrasound with software alignment. *Journal of Speech, Language and Hearing Research* .

Ohala, John J. 1981. The listener as a source of sound change. In *CLS 17: papers from the parasession on language and behavior*, ed. C.S. Masek, R.A. Hendrik, and M.F. Miller, 178–203. Chicago: CLS.

Ratelle, Georgia. 2009. The phonological patterning of glottal consonants. University of Ottawa MA mémoire.

Zharkova, Natalia, and Nigel Hewlett. 2008. Measuring lingual coarticulation from mid-sagittal tongue contours: Description and example calculations using english /t/ and /a/. *Journal of Phonetics* 37:248–256.