RECONSTRUCTING THE SOUNDS OF WORDS FROM THE PAST

John Coleman[1], John Aston[2], and Davide Pigole[2]

[1]University of Oxford Phonetics Laboratory, [2]University of Cambridge Statistical Laboratory
john.coleman@phon.ox.ac.uk, jada2@statslab.cam.ac.uk dp497@cam.ac.uk

ABSTRACT

We are developing novel statistical and signal processing methods to work backwards from contemporary audio recordings of simple words in modern Indo-European languages to regenerate audible spoken forms from earlier points in the language family tree. In this paper we present our first tentative steps in developing some of the necessary technical methods for realising this ambition, especially audio reconstruction of sound changes relating spoken Latin to French, Italian, Spanish and Portuguese.

Keywords: Phylogenetic, acoustic, historical

1. INTRODUCTION

Since the 19th century, philologists have studied evidence of sound change to infer the forms of words from a time before writing. For example, from Old English *weorc*, Latin *orgia*, Greek *ergon*, and Armenian *gorc*, philologists infer a Proto-Indo-European stem *u̯erĝ-*, hinting at a pronunciation something like [werg]. But what did it *actually* sound like? We are developing novel computational methods to regenerate *audible, historic, spoken forms* from contemporary audio recordings of simple words in modern Indo-European languages. These experiments open up many new questions: How far back in time can extrapolation from contemporary recordings reach? How diverse must a language family be in order to triangulate to sounds that are consistent with written forms from antiquity? How well do trees fit patterns of variation in the data (cf. [11])? Are any attested sound changes beyond the limits of the acoustic transformations we can currently model, and if so, how to address that? Though we hardly begin to answer such questions, in this paper we report on our initial steps in historical acoustic phonetics.

The relevance of quantitative biological models to the study of language change is generally accepted [6, 8] and has led to some impressive empirical results [3], yet such studies continue the philological practice of studying the history of words through their *written* forms. In programmatic work, the mathematical foundations have been laid for modelling the evolution of continuous functions, which can represent aspects of speech such as articulatory movements or acoustic features ([1, 10]). The present paper applies similar methods to modelling changes in a small test-vocabulary: the words for “one” to “ten” in six Romance varieties: I(talian), F(rench), (Castilian) S(panish), A(merican Spanish), P(ortuguese) and B(razilian Portuguese).

Differences between related languages may be traced back to dialect differences in a parent language, and to sound changes that applied in one dialect but not another (Figure 1). We do not consider words in living languages as equally modern; some embody more conservative, older pronunciation features. By “natural” variation and change we mean processes such as articulatory overshoot or undershoot (Figure 2), or changes in the phasing of articulators. Complete “deletion” or loss of sounds may arise as the end point of a chain of undershoot changes, e.g. Turkish *[ɑgɑtʃ] → [ɑɣɑtʃ] → [ɑɑtʃ]. Such explanations of the phonetic grounding of sound change are appealing [9, 2] but are *post hoc* and abstract.

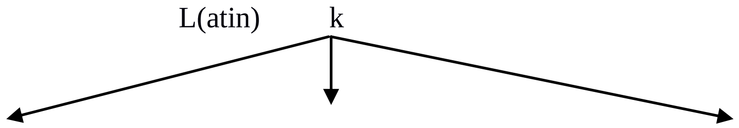
2. ACOUSTIC MODELLING OF SOUND CHANGES USING SPECTRAL MORPHING

2.1. Methods

Explanations framed in terms of articulatory parameters are difficult to test in practice, as articulatory measurement is often difficult and invasive. We therefore model processes of sound change in the *acoustic* domain, as there exist excellent tools for acoustic analysis, transformation and synthesis of audible sounds. The main techniques and resources for the acoustic reconstruction of spoken words are: 1) sound recordings of the words in question, spoken by many speakers in different languages; 2) extraction of acoustic parameters suitable for further numerical transformations; 3) computation of distance metrics and interpolants (morphing) to generate continua of forms in between recorded forms; 4) if possible, extrapolation beyond the ends of such continua to generate previously unheard forms, of greater antiquity; 5) re-synthesis of transformed parameters back into audible speech-like signals.

An acoustic simulation of sound change from *x* to

Figure 1: Incremental phonetic changes in Latin *centum*, *cantare*, *causa*, and *corpus* give rise to patterns of language differences conventionally modelled as a tree.



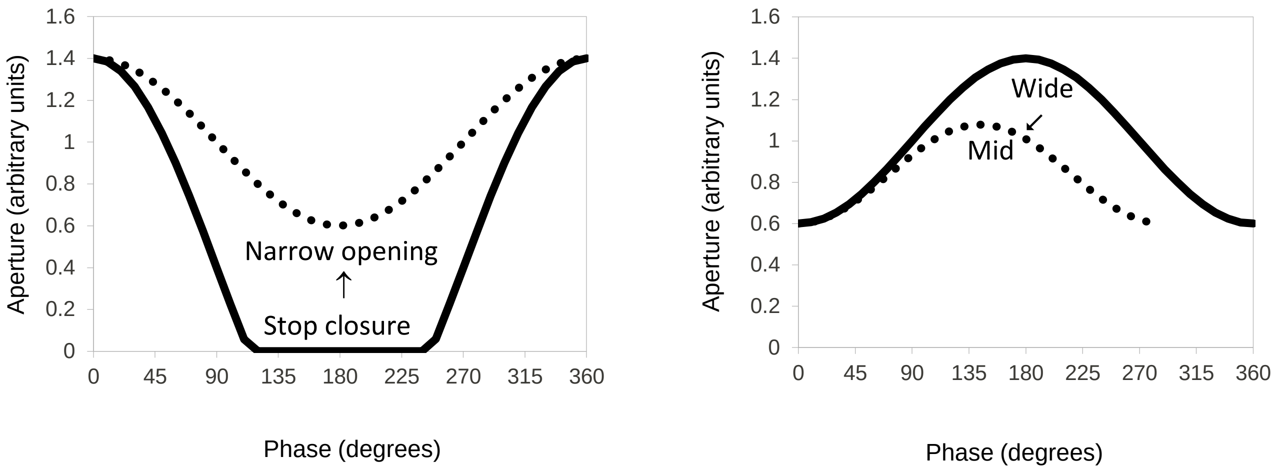
Velar		k̟ / _ i, e	L. [kentum]	k / _ a	L. [kantare], [kausa]	k / _ o, u ['korpus]
↓		↓	*[centu]	↓		
Palatal		cç	*[cçentu]	*kj		
↓		↓		↓		
Postalveolar	e.g. I.	tʃ	I. [tʃɛnto]	*cç	*[cçantare]	I. ['kɔrpo]
↓		↓		↓		
Alveolar		ts	*[tsẽnt]	*tʃ	Old F. [tʃãnter], [tʃɔzə]	
↓		↓	*[sãnt]	↓		
e.g. Older S, Mod. F.		s	*[sjen], F. [sã]	ʃ	F. [ʃãte], [ʃɔzə]	F. [kɔʀ]
↓		↓				
Dental	S.	θ	S. [θjen]			S. ['kwerpo]

y can be made if an audio recording *x'*, similar to the supposed historical form *x*, can be obtained. This may be from a living language displaying a more conservative pronunciation that has not undergone the sound change in question). For example, to model the development from L. *unus/unum* to P. *um* or F. *un*, though original audio recordings of L. are of course unavailable, the relevant segment of the original L., *un-*, can be simulated using portions of recordings as proxies for their likely pronunciation in L. Thus, taking the [u:n] sound of I. or S. *uno* as a plausible proxy for the sound of *un-* in L. *ūnus*[1], we model the acoustic-historical path from L. [u:n-] via P. [ũ] (Figure 3,) to F. [œ̃] and thence to the more recent F. pronunciation [ɛ̃] .

Audio recordings of isolation-form pronunciations of words for the digits “one” to “ten” in six Romance varieties (I., F., S., A., P. and B.) were obtained from the internet sites such as language classes. All recordings were converted using *sox* from their original formats (e.g. MP3) to monophonic 16-bit PCM files with a sample rate of 11,025 s^{-1}. Leading and trailing silences were trimmed so that tokens of a given word were optimally time-aligned, but no other timing characteristics were manipulated, i.e. there was no

Figure 2: a) Undershooting a closing movement; complete closure (*solid line*) → close approximation (*dashed line*) e.g. t → s, d → ð.

b) Undershooting an opening movement; large opening (*solid line*) → smaller opening over shorter time interval (*dashed line*) e.g. a → ə.



time warping. Voicing, f_0 and noise source parameters were estimated in 5 ms frames using the ESPS *get_f0* function ([4]), and 17th-order LPC reflection coefficients were derived by the Burg method, using the ESPS *refcof* function. The choice of LPC coding was driven by the ease with which such parameters can be automatically derived from speech recordings and converted back into quite natural-sounding synthetic speech.[2] The matrix of source and spectral parameters is a faithful acoustic representation of the original sound recordings, and is amenable to digital transformations such as morphing, following the methods of [7]. For selected pairs of recorded words, 9 intermediate source+spectral parameter matrices were interpolated using Matlab, to yield 11-step continua of forms ranging from 100% language 1 (0% language 2) to 0% language 1 (100% language 2). Sound files were synthesised from these parameter matrices using the ESPS *lp_syn* function. In this way we generated acoustic continua modelling a variety of historical processes in the development of Romance digits (Figure 4).

2.2 Results

The acoustic-historical path from Latin [u:n-] to the more recent F. pronunciation [ɛ̃] exemplifies a combination of several historical changes: (a) nasalisation of [u:] before [n], (b) loss of [n], yielding [ũ], as in P.; (c) lowering of the vowel to more open qualities, as in F. [œ̃] and more recently [ɛ̃]. In traditional symbolic phonology (a) might be thought to precede (b), because the nasalisation of the vowel would have no explanation if the nasal consonant were lost first; whereas vowel nasalisation introduces redundancy which makes the loss of the consonant less problematic. In our synthetic continuum of sounds from "Latin" [u:n-] to P. [ũ], however, the increasing nasalisation of the vowel and the loss or vocalisation of the nasal progress together. Interestingly, in the continuum from P. [ũ] to F. [ɛ̃], the older and more conservative F. form [œ̃] is derived as an intermediate point along the continuum, i.e. an outcome of the theoretical apparatus that agrees with observed data. The continuum from I. [due] to F. [dœ] illustrates monophthongisation, and that from I. [tre] to F. [tʀwa] illustrates diphthongisation (with change in the acoustics and articulation of [r]). In this case, the modelled development of F. [tʀwa] is not historically accurate; earlier forms such as*[trois] are suggested by the spelling, but do not "fall out" of our interpolation method. Monophthongisation is also illustrated by the continuum from S. [seis] to F. [sis], and diphthongisation by S. [dos] to B. [dois].

Figure 3: Acoustic modelling of sound change using speech coding and interpolation.

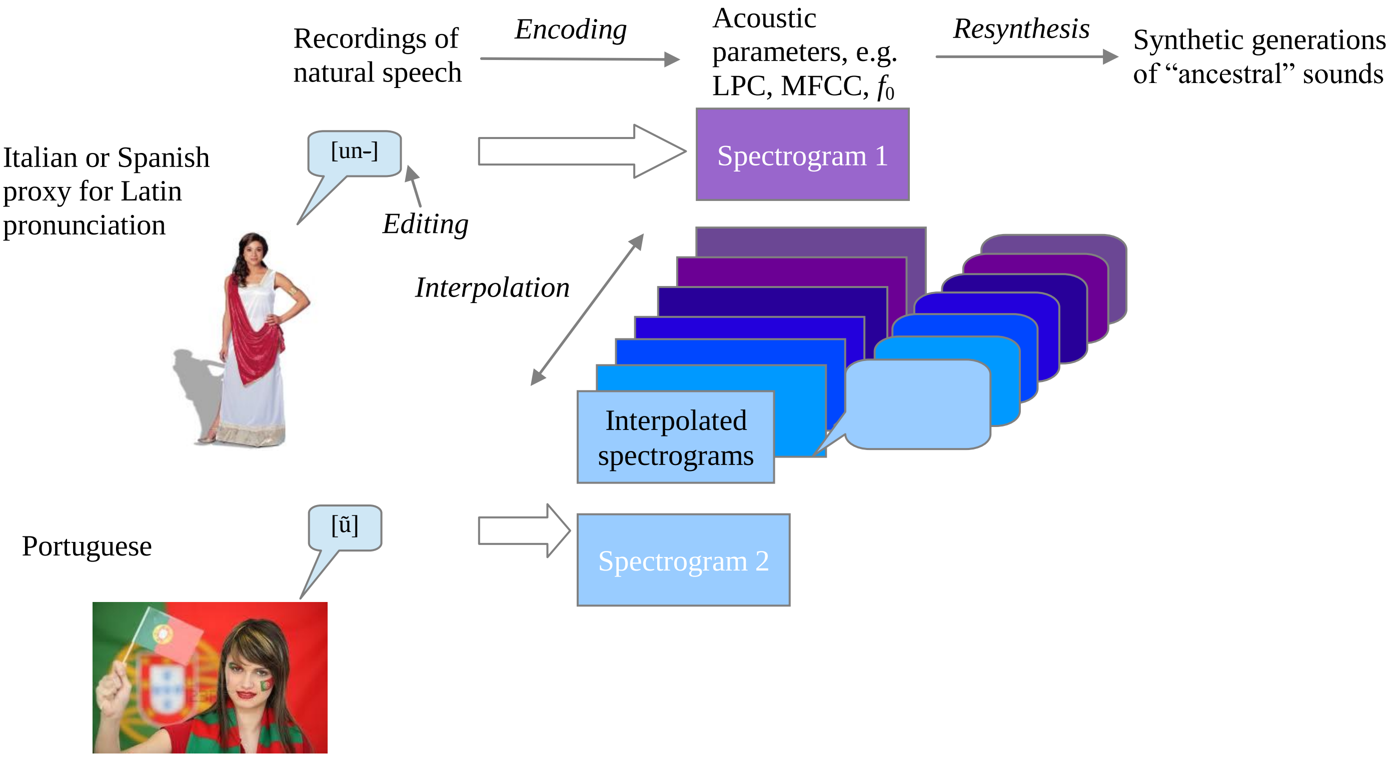

Recordings of natural speech
Encoding
Acoustic parameters, e.g. LPC, MFCC, f_0
Resynthesis
Synthetic generations of "ancestral" sounds
Italian or Spanish proxy for Latin pronunciation
[un-]
Editing
Spectrogram 1
Interpolation
Interpolated spectrograms
Portuguese
[ũ]
Spectrogram 2

Figure 4: Sound changes from the history of Romance that are modelled in this paper.

1. UNUS	2. DUO		3. TRES		6. SEX		7. SEPTEM	8. OCTO	10. DECEM
I. [un-]	I. [due]	S. [dos]	S. [tres]					I. [ot:o]	
↓	↓	↓	↙	↘				↓	
P. [ũ]	↓	B. [dois]	↓	I. [tre]	S. [seis]		F. [set]	↓	B. [des]
↓	↓	↓	↓	↓	↙	↘	↓	↓	↓
F. [œ̃]	F. [dœ]	P. [doiʃ]	P. [treiʃ]	↓	P. [seiʃ]	F. [sis]	P. [setʃ]	S. [otʃo]	F. [dis]
↓				↓					
F. [ɛ̃]				F. [tʀwa]					

The development from final [-s] to [-ʃ] after [i] in (Standard) P. is illustrated by B. [dois] → P. [doiʃ], S. [tres] → P. [treiʃ] , S. [seis] → P. [seiʃ] . Postalveolarization plus affrication is also seen in other dialects and contexts, e.g. F. [set] → P. [setʃ] and I. [ot:o] → S. [otʃo] .

3. FROM PRESENT-DAY RECORDINGS TO INFERRED ANCESTRAL SOUND FILES

The sound-morphing method employed in the demonstrations given above is simple linear[3] interpolation between a supposed ancestral form *A* and a modern recording *M*:

(1) $M = A + \mathrm{k}\,\delta_g$

in which k is the number of generations and δ_g the quantum of change per generation (both the magnitude of change and its direction). In our simulations, *M* and *A* are known, as “ancestral” sound recordings are proxies from a living language. Estimating each generation as 20 years, there are about 100 generations in the Romance tree, so a fine-grained history of word forms may in principle be derived. When *A* is known (a proxy), δ_g is simply $(M - A)/\mathrm{k}$. However, it is unusual for the ancestral forms of a language family to be as well documented as Latin; in general we wish to *infer A* on the basis of attested forms *M* and a theory of sound change providing a methodology for estimating δ_g.

To better understand δ_g, consider the derivation of P. [ũ] and S. [un-] from a common ancestor *L* *[un-]: $P = L + \mathrm{k}\,\delta_{gp}$ and $S = L + \mathrm{k}\,\delta_{gs}$. If *L* were not known, we might suppose that both modern languages have drifted away from Latin at approximately the same rate, but in symmetrically opposite directions $\delta_{gp} \approx -\delta_{gs}$. From this, we would infer a common ancestor that is a kind of average mid-way between *S* and *P*. But that is clearly not so: since *L* *[un-] is thought to be similar to both observed S. [un-] and I. [un-], $\delta_{gs} \approx 0$ and, since the number of generations is the same for both branches, most of the historical change is due to δ_{gp}. Inferring δ_g for each branch is helped by knowing which modern forms are more conservative and which are more innovating; this can be estimated to some extent from acoustic distances between modern recordings, in which the most innovative forms will be more distant from the centroid. For further work on distances and transformations which can be used to facilitate non-linear interpolation, see [12].

Where an ancestral form is not available, we may consider the modern forms in each daughter language as conservative proxies for the ancestral form, to calculate the innovations for as many different histories as there are daughter languages. To test these inferred histories, some of the calculated intermediate forms might be consistent with such written records of earlier forms as may be available, at an appropriate time-depth. For example, even if no Latin texts had ever remained, the spelling of words in Old F. (55–30 generations ago) provides some clues as to δ_{gf}; for example the “oi” and “s” of *trois* suggests an older pronunciation with final [s], as still seen in S. and B., and a diphthong that is phonetically intermediate between [ei] (as in *[treis], from L. *tres*) and French [wa] (in [tʀwa]).

The rate of change of languages (i.e. $|\delta_g|$) is not constant: as a result of historical events such as conquests, migrations, borrowing, etc. changes in one branch may be greater than in others and greater in some generations. Greater changes lead to splits in the family tree, whereas in less eventful times the pace of change can be very slow, as we saw in $\delta_{gs} \approx 0$ for *L* *[un-] ≈ *S* [un-] ≈ *I* [un-], a kind of “punctuated equilibrium” in historical development. However, methods and models in this approach can be extended to time-varying δ_g.

4. REFERENCES

[1] Aston, J. et al. (The Functional Phylogenies Group.) 2011. Phylogenetic inference for function-valued traits: speech sound evolution. *Trends in Ecology and Evolution* 27 (3), 160–166.

[2] Blevins, J. 2004. *Evolutionary Phonology.* Cambridge University Press: Cambridge.

[3] Bouchard-Côté, A., Hall, D., Griffiths, T. L., Klein, D. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proc. Nat. Acad. Sci.* 110 (11), 4224–4229.

[4] Entropic Research Laboratory, Inc. Entropic Signal Processing System. http://www.phon.ox.ac.uk/releases

[5] Erro, D., Sainz, I., Navas, E., Hernaez, I. 2011. Improved HNM-based Vocoder for Statistical Synthesizers. *Proc. InterSpeech*, Florence, 1809–1812.

[6] McMahon, A., McMahon, R. 2003. Finding families: quantitative methods in language classification. *Trans. Philological Soc.* 101 (1), 7–55.

[7] Moore, D., Coleman, J. 2005. Generation of synthetic speech. US Patent Application 20050171777.

[8] Nakhleh, L., Warnow, T., Ringe, D., Evans, S. N. 2005. A comparison of phylogenetic reconstruction methods on an Indo-European dataset. *Trans. Philological Soc.* 103 (2), 171–192.

[9] Ohala, J. J. 1993. The phonetics of sound change. In: Jones, C. (ed), *Historical Linguistics: Problems and Perspectives*. Longman: London. 237–278.

[10] Pigoli, D., Aston, J.A.D., Dryden, I. L., Secchi, P. 2014. Distances and Inference for Covariance Operators. *Biometrika* 101, 409–422.

[11] Shiers, N., Aston, J. A. D., Smith, J. Q., Coleman, J. Gaussian tree constraints applied to acoustic linguistic functional data. http://archiv.org/abs/1410.0813

[12] Pigoli, D., Hadjipantelis, P., Coleman, J., Aston, J. A. D. The analysis of acoustic phonetic data: exploring changes in the spoken digits of Romance languages. Under review.

[1] Audio clips in MP3 format embedded in this document may be heard if your PDF reader enables it. These audio examples are also available in .wav format at http://www.phon.ox.ac.uk/jcoleman/ancient-sounds-presentations.html.

[2] We are in the process of replicating this experiment using MFCC-based analysis-synthesis software, *ahocoder* and *ahodecoder* [5]. This promises to yield more natural-sounding synthetic interpolants than LPC resynthesis. Another possibility is to use Fourier time-frequency spectrograms, as illustrated in [12].

[3] Of course the true path of development might not be linear, but as *A* and *M* are acoustic representations of many (e.g. 20 to 40) dimensions we don't wish to complicate the model with nonlinear terms yet (but see [12]).