

A topological definition of voice mimickers

Amedeo De Dominicis

Università della Tuscia, Viterbo, Italy <dedomini@unitus.it>

This paper applies a topological model to analyze the degree of similarity between formant charts. The formant charts of the stressed vowels of some speakers of some voice mimickers (computed on the basis of the mean values of F_1 and F_2) represent the initial data. The model generates predictions that have been compared and validated by a perception test.

KEYWORDS: topology, phonetics, voice mimickers.

1. Introduction

This paper concerns the analysis of voice mimickers and implements the topological model elaborated in De Dominicis (2020). It deals with some impersonations primarily designed for entertainment purposes in radio broadcasts, but is not primarily about impersonation as such. It is a ‘method study’, since its main goal is not to evaluate how successfully the impersonators managed their task, but to test if a given topological morphing method can be used to quantify the degree of success of a given imitation.

The method used in the acoustic analysis is a form of topological transformation between the vowel spaces representing imitators, imitation, and targets. Based on these transformations, similarity indices between the compared vowel spaces have been calculated.

We investigate only the formant frequencies of the stressed vowels. We do not analyze the timing (articulation rate of phones and words) or the fundamental frequency.

The model allows us to analyze the degree of similarity between formant charts. The formant charts of the stressed vowels of a number of speakers (target and imitated voices) represent the initial data. The model generates predictions about their topological equivalence. These similarity predictions would, however, be of limited interest in linguistics were they not shown to correlate with human similarity perception. To address this question, a perception test was included in the study.

2. State of the art and previous works

There are organic differences between speakers, which cannot be changed, making it difficult to produce exact copies of another speaker's voice and speech. Despite these differences, generally speaking, we assume that every imitator manages to use his vocal setting in order to overcome his natural limits and succeed in imitation. We are interested in evaluating the latter and not the former, as we focus on the impersonator's phonetic skills and normalize all organic individual differences. To get close to the target speaker and to succeed in voice imitation, the impersonator must change his own voice and speech behavior in a number of ways. He must identify important and characteristic features of the target speaker's voice and speech style and know how to change his own voice in order to succeed with the impersonation.

Formant frequencies are one of the acoustic correlates of differences between vowels and are primarily determined by the shape characteristics of the speaker's vocal tract. They are identified by the peaks in the spectral envelope of the speech signal and determined by the natural resonances of the vocal tract. For a given speaker, changes in formant frequencies depend primarily on changes in the shape and position of the articulators (tongue, lips, jaw, etc.), according to the so-called source-filter theory developed by Gunnar Fant (Fant 1960, 1966).

Given its methodological nature, this study does not rely on previous references (except for De Dominicis 2020). Nonetheless, we concisely summarize some findings of the specialistic literature. Previous studies on vocal mimicry have considered a multiplicity of factors, such as timing (at the segmental level), mean of formant frequencies and pitch of imitator and his target voice.¹ Bessler (1991) studied a caricatured impersonation of Charles de Gaulle. The most relevant finding with respect to that study was the fact that the impersonator exaggerated both mean fundamental frequency level and range.

In another study (Endres *et al.* 1971), vowel formant frequencies and fundamental frequencies in imitations were compared with the corresponding values of the original voices. Although the imitators managed to change their formant and fundamental frequencies in the direction of the target values, "they were not able to adapt these parameters to match or even be similar to those of the imitated persons" (Endres *et al.* 1971: 1842).²

Another study, by Eriksson & Wretling (1997), found that a professional impersonation artist, imitating three well-known Swedish

public figures, was able to mimic the global speech rate very closely, but timing at the segmental level showed little or no change in the direction of the targets. Mean fundamental frequency and variation matched the targets very closely. Target formant frequencies were attained with varying success. With respect to individual vowels it was generally, but not always, the case that the formant frequencies of the mimicked vowels were closer to the original than those of the artist's own voice. Word durations in the imitations correlate better with the impersonator's natural versions than with the targets. A study of timing at the segmental level confirmed the greater similarity between the natural versions and the imitations than between the imitations and the targets. Articulatory timing at the segmental level changed very little in the imitations. The results obtained by Eriksson & Wretling (1997) indicate that timing at the segmental level is the least flexible aspect of speech production, that is, the aspect a speaker is least able to modify in a desired direction.

Similarly, Wretling & Eriksson (1998) demonstrated that timing patterns in speech are fairly stable within a speaker. Their findings, based on phoneme level data, agree well in this respect with the results obtained on word level data for the same speech material (Eriksson & Wretling 1997).³

Mejvaldová (2004) found that the timing properties (global duration and segmental durations) are the most stable characteristics of the speaker and cannot be easily imitated, whereas F0 is more imitable.⁴

The flexibility of just pitch and formant frequencies was confirmed by Kitamura (2008), who conducted a comparative study of a voice produced by a professional impersonator imitating a target speaker. Comparison of pitch frequency showed that the mean and dynamics of the pitch frequency of the imitated voice are changed so that they become closer to those of the target voice. The spectra of vowels uttered by the speakers are also similar in their shape and formant frequencies. In the imitated voice, the formant frequencies shift by up to 68% from those of the impersonator's natural voice.

The same results were obtained by Hautamäki *et al.* (2015) in a study of two impersonators imitating the voice of eight well-known Finnish public figures: they were able to adapt the fundamental frequency (F0) especially, but only occasionally the formant frequencies, towards the target speakers (cf. also Farrús *et al.* 2010, Perrot *et al.* 2007, Zetterholm 2007).

3. The model

As noted earlier, the model was elaborated in De Dominicis (2020: 62-74) and we refer readers to §2 and §3 of that paper, where one can find a detailed description of the mapping of formant charts and of the methodological specifications of the model. A summary description follows.

A formant chart is a schematic plotting of the vowels' formant frequencies and represents the vowels of a speaker. The vertical axis of the diagram denotes vowel height, with high vowels at the top of the diagram, whereas the horizontal axis indicates the anterior/posterior space, with front vowels to the left of the diagram. The vertical axis represents the values (in Hz) of the first formant F_1 (in reverse order, i.e. with lower values corresponding to high vowels and vice versa) while the horizontal axis represents the values of the second formant F_2 (with higher values corresponding to front vowels and vice versa). The first two formants are important in determining the quality of vowels and are frequently said to correspond to the open/closed and front/back dimensions. A formant chart represents the vocal tract, i.e. the vocal setting of a speaker: a geometrical representation, based on a bidimensional polygon (e.g. a trapezoid).⁵ We adopt the Hz scale even though the Bark scale would be a better predictor in acoustic terms. This choice is discussed in De Dominicis (2020: 62-63; 95, note 1; and an empirical comparison is illustrated in Appendix 3, pp. 7-10).⁶

The topological model is used to build a metrics relating different formant charts to one another and to calculate their rate of similarity. The method allows one to compare – holistically – the shapes of two (or more) formant charts. The advantage comes from comparing global vocal sets rather than single points or isolated vowels.

The technical tool adopted for implementing this model is the Mac application *D'Arcy Thompson's Pictures* (or **DTP** – <www-groups.dcs.st-and.ac.uk/~john/darcy.html>)⁷ (cf. Thompson 1917). Such an application enables users to alter a geometrical figure in real time by varying parameters in mathematical functions. DTP uses quadratic maps, that is, maps of the form $f(x,y) = (p(x,y), q(x,y))$ where p and q are polynomials of degree 2 in two variables. Hence the user has the freedom to vary 10 parameters. With so many degrees of freedom, quadratic maps allow one to vary the parameters (the values of x and y) continuously and to follow the results of such variations.

Since each formant chart is a geometrical figure, DTP enables one to calculate the topological transformation that occurs from one given

formant chart to another (say, from A to B). We call this transformation ‘actual-T’.

In DTP the function that accounts for the absence of a topological transformation of a given polygon (say, from A to A) is defined as $((0x^2 + 0xy + 0y^2 + 1x + 0y), (0x^2 + 0xy + 0y^2 - 0x + 1y))$. We call it ‘default-T’ or null-T.

By comparing default-T and actual-T, we obtain a numerical measure of what we call the Similarity Index (SI). SI is defined as follows: $SI = \Sigma (\Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 + \Sigma\Delta_5 + \Sigma\Delta_6)$. Where

- Δ_1 (x-range) = Δ between the value of x-range in actual-T and in default-T ($-2.5 \leq x \leq 0$)
- Δ_2 (x-range) = Δ between the value of x-range in actual-T and in default-T ($0 \leq x \leq +2.5$)
- Δ_3 (y-range) = Δ between the value of y-range in actual-T and in default-T ($-2.5 \leq y \leq 0$)
- Δ_4 (y-range) = Δ between the value of y-range in actual-T and in default-T ($0 \leq y \leq +2.5$)
- Δ_5 = Δ between the value of each of the five variables in actual-T and in default-T ($0x^2 + 0xy + 0y^2 + 1x + 0y$)
- Δ_6 = Δ between each of the five variables in actual-T and in default-T ($0x^2 + 0xy + 0y^2 + 0x + 1y$).

The closer the SI value is to 0, the more similar the polygons.

Given the polygons A and B, DTP enables one to calculate a $T(A) = B'$ and an inverse- $T(B) = A'$. But in order to avoid any potential gap between B and B', and between A and A', here and in De Dominicis (2020: 68-69) we adopt a ‘rejection statement’, which claims that if B and B' (or A and A') are patently different,⁸ then they are not considered as belonging to the same topological class. This statement avoids any further measurement between two figures that are not patently equal. If the rejection statement does not apply (i.e. if T does generate a figure more or less similar to what is expected), then we must obtain a measure of that similarity and we do so by applying the measurement of SI. For details and explanations about the rejection statement, we refer readers to §2 and §3 of De Dominicis (2020: 62-74).

In the case of impersonation, the implementation of the topological model differs from the one in De Dominicis (2020). Actually, in the case of impersonation we must measure the vocal distance, that is, the gap the impersonator must fill in order to successfully imitate his target, whereas in De Dominicis (2020) the purpose is to compare the similarity rate between two voices, speakers, or dialects. In fact, in imitation we presume that the original voice of the impersonator and of his target are not similar, but become similar because the impersonator carries out vocal ‘work’ to accomplish a tuning to his target. In De Dominicis (2020) the SI is the value used to measure the similarity rate between voices, and the smaller the SI, the greater the similarity rate. Conversely, here

the quality of the impersonator's 'phonetic work' is measured by comparing the arithmetic means between all T, as explained in the following section.

4. Implementation of topology on phonetics

The acoustic analysis used in this study must be interpreted within a linguistic context. In order to assign the SIs of T to assess the identity of two formant chart polygons (and the vowels they represent), an appropriate method must be elaborated. The transformations on polygons representing formant spaces are very special figures. In fact, differently from other figures submitted to a topological transformation, they have no 'holes': thus, their transformations are not subordinated to the restriction of the topological theory concerning the invariance of the number of holes between the original and the transformed figure.

Moreover, they are special because inversion by symmetry is not allowed. For phonetic reasons, this transformation is illegitimate in the case of polygons representing formant spaces: an [u] will always be on the right of [i], and an [a] always below [u] and [i].

Given A as the polygon representing the imitator's formant space, B as the polygon representing his target (the voice of the imitated person), and C as the polygon representing the formant space of his imitation, then T_1 is the topological transformation from A to C, and T_2 is the topological transformation from B to C. Now, provided there is no rejection, due to the implementation of our rejection statement, T_1 is defined by its SI (SI_1) and thus T_2 by its SI_2 .

The experimental expectation is that if the values of SI_1 and SI_2 are very similar, then A and B are similar and they are called 'compatible'. Nevertheless, this evaluation must be normalized, i.e. measurable in relation to A and B, that is to the original voices of the imitator and of his target. Otherwise, the similarity between SI_1 and SI_2 could be very high in the case where A and B are very similar and vice versa. Thus, we also take into account T_3 (the transformation from A to B) and its SI_3 , and T_4 (the transformation from B to A) and its SI_4 . We then calculate the arithmetic mean of SI_1 and SI_2 $((SI_1 + SI_2)/2)$ and of SI_3 and SI_4 $((SI_3 + SI_4)/2)$. We call **SImM** (mean SI between Mimicking voice and original voices) the first mean (between SI_1 and SI_2) and **SImOV** (mean SI between the Original Voices of the impersonator and his target) the second mean (between SI_3 and SI_4). If $SImM \leq SImOV$, then A and B are very compatible voices; if $SImM > SImOV$, then A and B are not compatible voices.⁹ In order to calculate the degree of similarity

between compatible voices and determine a ranking of similarity among imitations, we use the quotient SI_{mOV}/SI_{mM} : the smaller the result, the greater the similarity rank and the better the imitation. We call this result **IR** (Imitation Rank).

From a phonetic standpoint, all four similarity measures (SI_{1-4}) are relevant and necessary. SI_{1-2} refer to the imitation compared to the imitator (SI_1) and to the target (SI_2). However, SI_{1-2} must be normalized: the phonetic quality of the imitation could be biased by organic factors (e.g. anatomical similarity/dissimilarity between imitator and target). SI_{3-4} provide this information because both refer to the anatomical bases of the speakers. In particular, the appropriate parameter is the relation (IR) between SI_{mOV} and SI_{mM} : it takes into account all four SIs and compares the phonetic factors (SI_{1-2}) to the possible organic similarity of the speakers (SI_{3-4}) and provides their normalization. The normalization aims to avoid any influence of extralinguistic and individual factors (due to a given imitator and a given target) from the linguistic evaluation of an imitation; SI_{mOV} and SI_{mM} result from an arithmetic mean, because they average the contribution of all SIs (SI_{1-4}).

If a given imitator and his target were anatomically ‘similar’, then the imitation would be easier, favored, and the imitator would be easily successful; if not, then the imitation would be anatomically biased, and – above all – the former and the latter imitation would be phonetically incomparable.

5. The voices of impersonators, their imitations, and their original targets.

We will analyze and compare the original voices of two famous Italian professional impersonators, their imitations of some well-known personalities (target voices), and also the original voices of these celebrities. The analysis will deal with the formant frequencies of their stressed vowels.¹⁰

The audio corpus was made available by RAI TECHE, the archive of the national public broadcasting Company of Italy. All recordings come from old radio broadcasts, and are in wav format, mono channel, sampling frequency 44,1 kHz. In all (four) cases the corpus is divided into three sub-corpora: the original voice of the impersonator, the original voice of the imitated target, and the voice of the impersonator imitating the target. For each sub-corpus we extracted the values in Hz of the vowel formants and plot them into a formant chart. The vowels were chosen in the same phonological context: stressed, and preceded by the

same consonant, or by a homorganic consonant (a consonant with the same place of articulation).¹¹

5.1. The case of Tortora-Califano

Max Tortora is a famous Italian professional impersonator. Franco Califano was an Italian singer who died in 2013. We analyzed a recording of both voices (the original voice by Tortora, Tortora imitating Califano, and the original voice by Califano) taken from the radio broadcast *Il cammello di Radio2-Picnic* on 21/7/2005.¹² In this case we analyzed only 3 vowels ([e], [a], [o]), because the recordings did not allow us to find instances of [i] and [u] in all sub-corpora.

Table 1 shows the formant frequencies of each sub-corpus.¹³

SUB-CORPORA VOICES	WORDS	VOWELS	F1	F2
Tortora-original	<u>m</u> e ('me')	e	419	1699
	qu <u>a</u> nno ('when')	a	614	1429
	ri <u>o</u> ('I go back')	o	524	1032
Califano target	<u>m</u> e ('me')	e	422	1947
	qu <u>a</u> rto ('forth')	a	689	1305
	st <u>o</u> ria ('history')	o	341	1281
Tortora imitating Califano	<u>m</u> e ('me')	e	372	1624
	qu <u>a</u> nno ('when')	a	631	1619
	to <u>o</u> ('I go back')	o	619	1022

Table 1. Formant frequencies of each Tortora-Califano sub-corpus. Stressed vowels are underscored.

Below, figures 1-5 show the formant charts and their T, which refer to each Tortora-Califano sub-corpus.

Figure 1 shows the corresponding formant charts of the voices of Tortora imitating Califano, of Califano-target, and of Tortora-original.

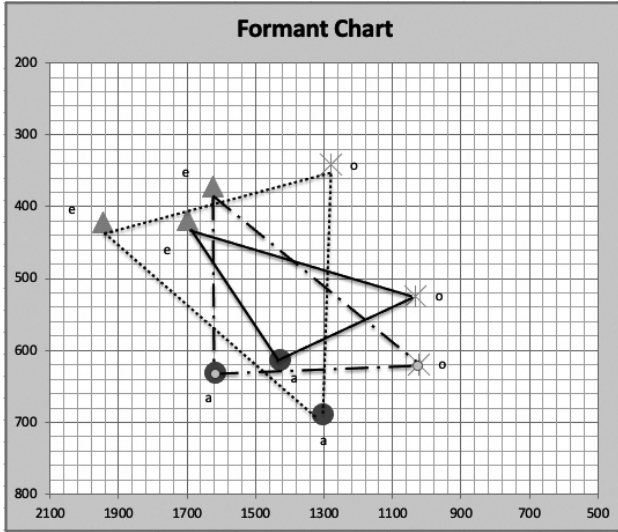


Figure 1. Formant chart of Tortora-Califano: the dash-dot line polygon refers to Tortora imitating Califano, the dotted line one refers to Califano target, and the solid line one refers to Tortora-original.

Figure 2 shows the T from the polygon of Tortora-original to the one of Tortora imitating Califano. The $SI_1 = 0.925$.

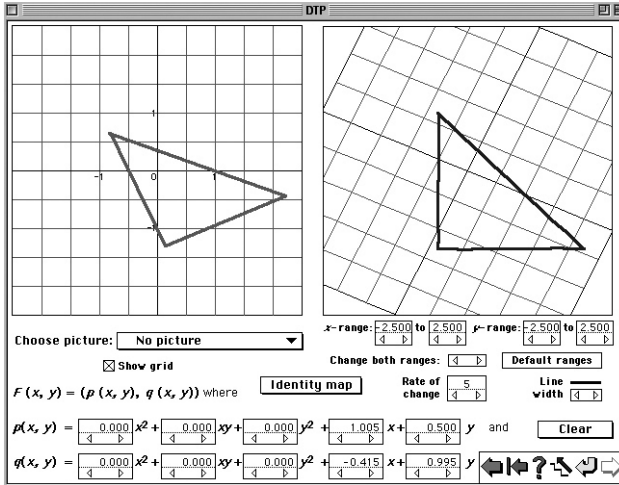


Figure 2. T from the polygon of Tortora-original (left) to the one of Tortora imitating Califano (right).

Figure 3 shows the T from the polygon of Califano target to the one of Tortora imitating Califano. The $SI_2 = 11.27$.

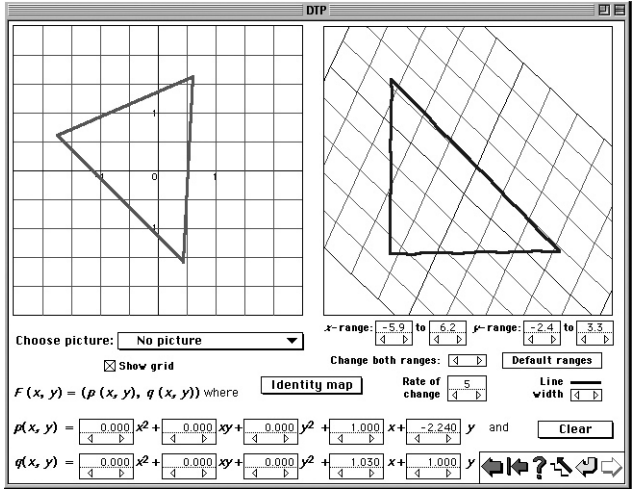


Figure 3. T from the polygon of Califano target (left) to the one of Tortora imitating Califano (right).

Figure 4 shows the T from the polygon of Tortora-original to the one of Califano target. The $SI_3 = 8.465$.

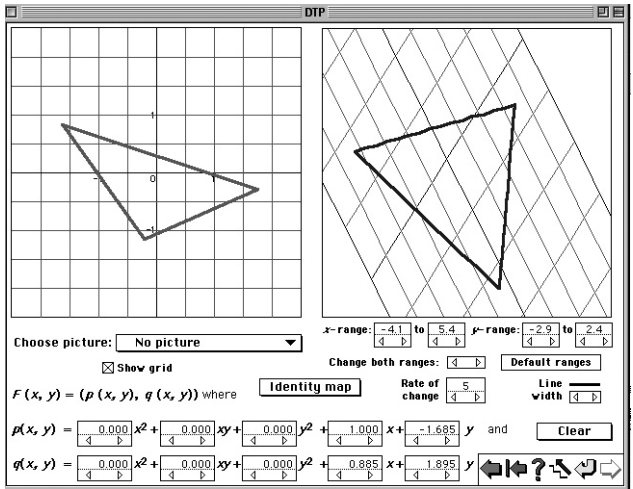


Figure 4. T from the polygon of Tortora-original (left) to the one of Califano target (right).

Figure 5 shows the inverse-T from the polygon of Califano target to the one of Tortora-original. The $SI_4 = 17.616$.

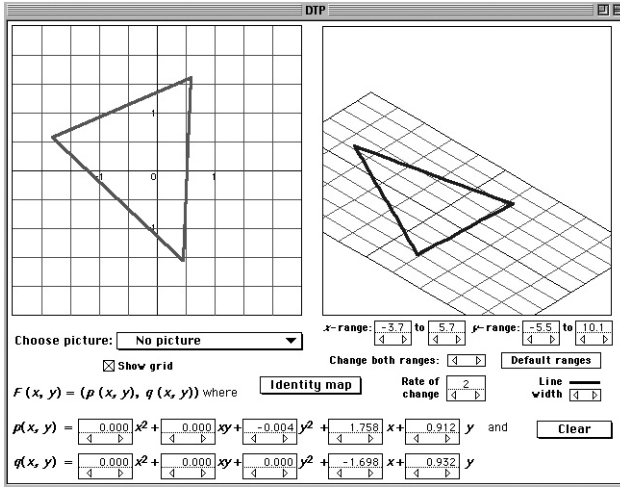


Figure 5. Inverse-T from the polygon of Califano target (left) to the one of Tortora-original (right).

The arithmetic mean of SI_1 and SI_2 (SimM) is: $(SI_1 + SI_2)/2 = 6.0975$. The arithmetic mean of SI_3 and SI_4 (SimOV) is: $(SI_3 + SI_4)/2 = 13.0405$. Thus, $SimM < SimOV$. As a consequence, the voices of Tortora and Califano are fully compatible. The IR of this imitation is 2.138.

5.2. The case of Tortora-Sordi

Max Tortora is the professional impersonator mentioned earlier. Alberto Sordi was a famed Italian actor who died in 2003. We analyzed a recording of Tortora imitating Sordi during the radio broadcast *Ottantaradio* on 4/10/2004;¹⁴ and a recording of the original voice by Sordi in the radio broadcast *Hollywood Party* on 18/4/2013. Tortora's imitation of Sordi is very trendy and popular in Italian media.

Table 2 shows the formant frequencies of each sub-corpus.

SUB-CORPORA VOICES	WORDS	VOWELS	F1	F2
Tortora-original	<i>v<u>i</u>ta</i> ('life')	i	263	2145
	<i>v<u>e</u>de</i> ('he sees')	e	385	2014
	<i>c<u>a</u>so</i> ('case')	a	695	1461
	<i>pot<u>u</u>to</i> ('could')	u	174	954
	<i>con<u>o</u>sco</i> ('I know')	o	460	865
Sordi target	<i>ri<u>v</u>ista</i> ('magazine')	i	318	2055
	<i>av<u>e</u>ssi</i> ('[if] I had')	e	464	1888
	<i>ca<u>r</u>a</i> ('dear' F.SG)	a	709	1465
	<i>pot<u>u</u>to</i> ('could')	u	385	1280
	<i>con<u>o</u>sco</i> ('I know')	o	430	1057
Tortora imitating Sordi	<i>v<u>i</u>ta</i> ('life')	i	354	1918
	<i>ve<u>d</u>i</i> ('you see')	e	374	1842
	<i>ca<u>s</u>a</i> ('house')	a	816	1488
	<i>tut<u>t</u>e</i> ('all' F.PL)	u	377	1222
	<i>con<u>o</u>sce</i> ('he knows')	o	385	1237

Table 2. Formant frequencies of each Tortora-Sordi sub-corpus. Stressed vowels are underscored.

Below, figures 6-10 show the formant charts and their T, which refer to each Tortora-Sordi sub-corpus.

Figure 6 shows the corresponding formant charts of the voices of Tortora imitating Sordi, of Sordi-target, and of Tortora-original.

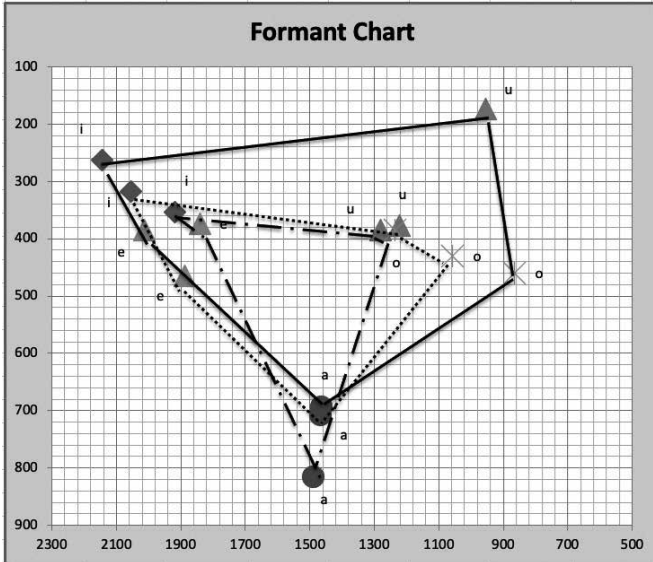


Figure 6. Formant chart of Tortora-Sordi: the dash-dot line polygon refers to Tortora imitating Sordi, the dotted line one refers to Sordi target, and the solid line one refers to Tortora-original.

Figure 7 shows the T from the polygon of Tortora-original to the one of Tortora imitating Sordi. The $SI_1 = 7.499$.

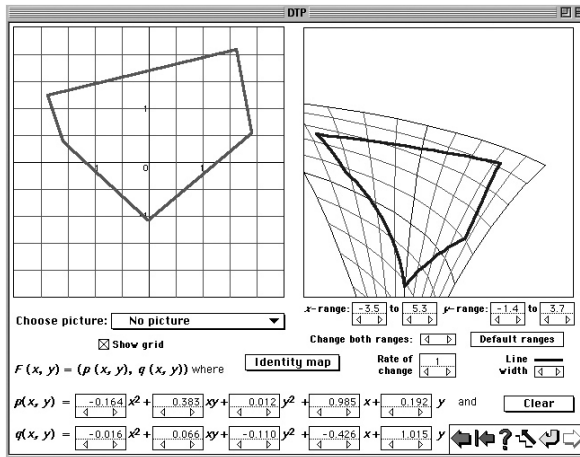


Figure 7. T from the polygon of Tortora-original (left) to the one of Tortora imitating Sordi (right).

Figure 8 shows the T from the polygon of Sordi target to the one of Tortora imitating Sordi. The $SI_2 = 5.724$.

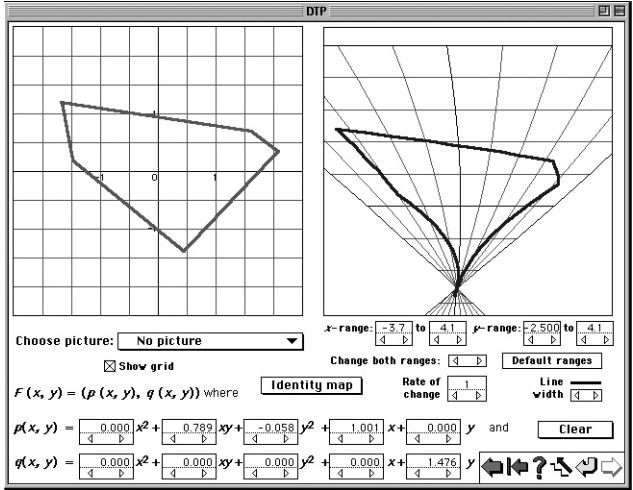


Figure 8. T from the polygon of Sordi target (left) to the one of Tortora imitating Sordi (right).

Figure 9 shows the T from the polygon of Tortora-original to the one of Sordi target. The $SI_3 = 34.2435$.

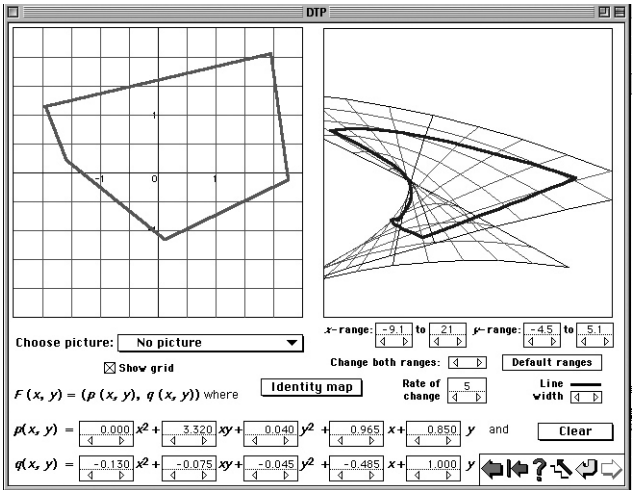


Figure 9. T from the polygon of Tortora-original (left) to the one of Sordi target (right).

Figure 10 shows the inverse-T from the polygon of Sordi target to the one of Tortora-original. The $SI_4 = 10.58$.

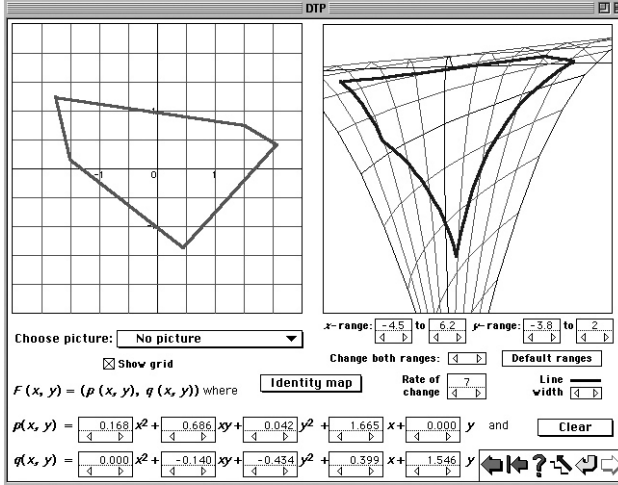


Figure 10. Inverse-T from the polygon of Sordi target (left) to the one of Tortora-original (right).

The arithmetic mean of SI_1 and SI_2 (SimM) is: $(SI_1 + SI_2)/2 = 6.6115$. The arithmetic mean of SI_3 and SI_4 (SimOV) is: $(SI_3 + SI_4)/2 = 22.41175$. Thus, $SimM < SimOV$. As a consequence, the voices of Tortora and Sordi are fully compatible. Moreover, the voices of Tortora and Califano are more similar than those of Tortora and Sordi since IR (Tortora-Califano) = 2.138, whereas IR (Tortora-Sordi) = 3.389. Thus, the imitation by Tortora is more compatible (and successful) with the target ‘Califano’ than with the target ‘Sordi’.

5.3. The case of Giusti-Lotito

Max Giusti is an Italian professional impersonator, and Claudio Lotito owns the ‘Lazio’ football team and is popular among Italian football supporters. We analyzed a recording of the original voice by Giusti, and Giusti imitating Lotito during the radio broadcast *Radio2 Super Max* on 20/5/2015;¹⁵ and a recording of the original voice by Lotito in the radio broadcast *GR1 Sport*, on 22/8/2013 at 19:30.

Table 3 shows the formant frequencies of each sub-corpus.

SUB-CORPORA VOICES	WORDS	VOWELS	F1	F2
Giusti-original	<i>ri<u>d</u>ere</i> ('to laugh')	i	169	2094
	<i>se<u>m</u>pre</i> ('always')	e	630	1507
	<i>fa<u>l</u>lo</i> ('he does')	a	640	1388
	<i>lu<u>i</u></i> ('he')	u	343	640
	<i>co<u>n</u>osce</i> ('he knows')	o	498	1234
Lotito target	<i>in<u>i</u>ziativa</i> ('initiative')	i	419	2017
	<i>qu<u>e</u>sti</i> ('these')	e	411	1949
	<i>ca<u>d</u>a</i> ('fall' subjunctive)	a	660	1364
	<i>u<u>n</u>a</i> ('one')	u	492	892
	<i>eco<u>n</u>omico</i> ('economic')	o	573	842
Giusti imitating Lotito	<i>se<u>n</u>tito</i> ('heard')	i	306	2139
	<i>qu<u>e</u>sto</i> ('this')	e	459	1661
	<i>ca<u>r</u>o</i> ('dear')	a	903	1502
	<i>pu<u>n</u>to</i> ('point')	u	380	954
	<i>no<u>n</u>no</i> ('grandfather')	o	423	862

Table 3. Formant frequencies of each Giusti-Lotito sub-corpus. Stressed vowels are under-scored.

Below, figures 11-15 show the formant charts and their T, which refer to each Giusti-Lotito sub-corpus.

Figure 11 shows the corresponding formant charts of the voices of Giusti imitating Lotito, of Lotito-target, and of Giusti-original.

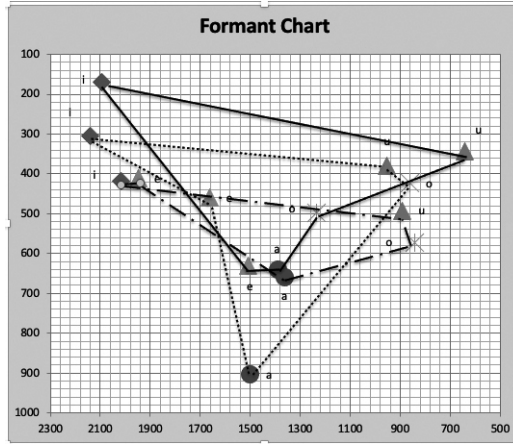


Figure 11. Formant chart of Giusti-Lotito: the dotted line polygon refers to Giusti imitating Lotito, the dash-dot line one refers to Lotito target, and the solid line one refers to Giusti-original.

Figure 12 shows the T from the polygon of Giusti-original to the one of Giusti imitating Lotito. The $SI_1 = 5.681$. But also in this case we must adopt our rejection statement: the polygon of Giusti imitating Lotito (on the right side of figure 12) is patently different from the polygon of Giusti imitating Lotito (in the dotted line of figure 11). Thus, they are not considered as belonging to the same class.

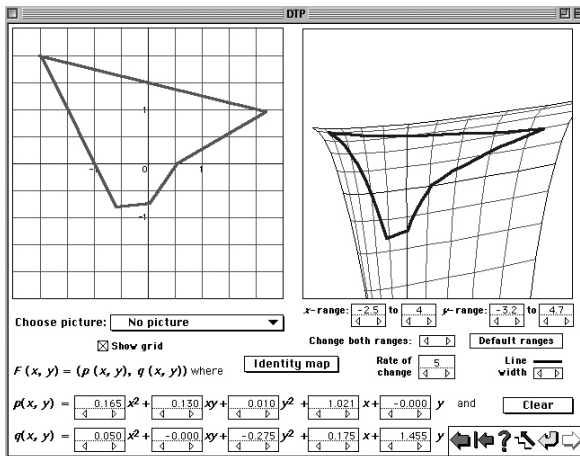


Figure 12. T from the polygon of Giusti-original (left) to the one of Giusti imitating Lotito (right).

Figure 13 shows the T from the polygon of Lotito target to the one of Giusti imitating Lotito. The $SI_2 = 4.343$. But in this case, we must adopt our rejection statement: if the polygon T(A) and the polygon B are patently different, then they are not considered as belonging to the same class. In fact, the polygon of Giusti imitating Lotito (on the right side of figure 13) is patently different from the polygon of Giusti imitating Lotito (in the dotted line of figure 11).

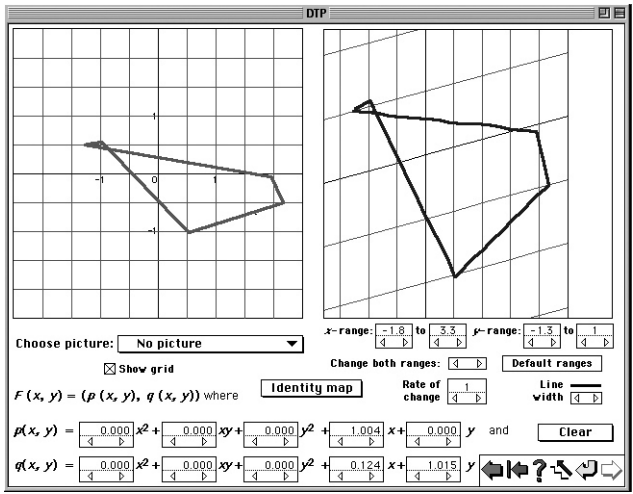


Figure 13. T from the polygon of Lotito target (left) to the one of Giusti imitating Lotito (right).

Figure 14 shows the T from the polygon of Giusti-original to the one of Lotito target. The $SI_3 = 21.5$. But in this case also we must adopt our rejection statement: if the polygon T(A) and the polygon B are patently different, then they are not considered as belonging to the same class. In fact, the polygon of Lotito target (on the right side of figure 14) is patently different from the polygon of Lotito target (in the dash-dot line of figure 11).

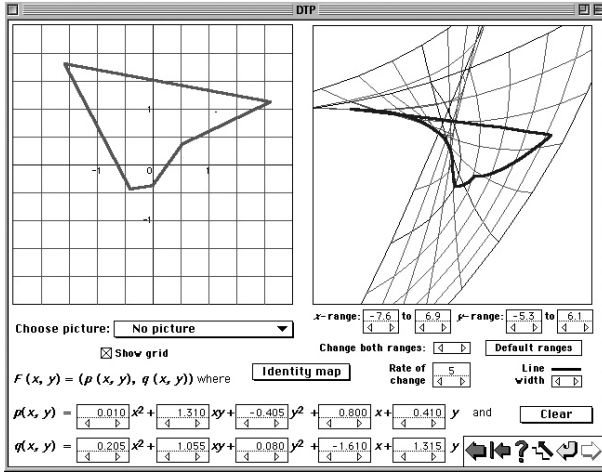


Figure 14. T from the polygon of Giusti-original (left) to the one of Lotito target (right).

Figure 15 shows the inverse-T from the polygon of Lotito target to the one of Giusti-original. The $SI_4 = 3.797$. But in this case also we must adopt our rejection statement: the polygon of Giusti-original (on the right side of figure 15) is patently different from the polygon of Giusti-original (in the solid line of figure 11). Thus, they are not considered as belonging to the same class.

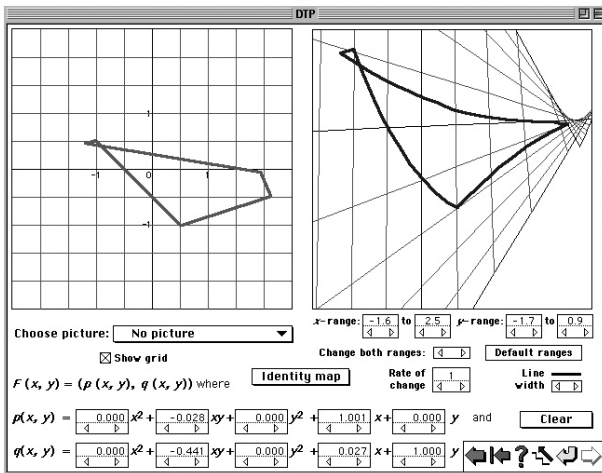


Figure 15. Inverse-T from the polygon of Lotito target (left) to the one of Giusti-original (right).

In the case of Giusti-Lotito the question about similarity is not appropriate. No T is able to pass the rejection statement, and thus their SI is negligible and not significant. Our representation predicts that the voices of the two speakers are not compatible, despite public appreciation.

5.4. The case of Giusti-Rossella

Max Giusti is the Italian professional impersonator mentioned earlier and Carlo Rossella is an Italian columnist relatively unfamiliar to the Italian public except for some newspaper readers. We analyzed a recording of Giusti imitating Rossella during the radio broadcast *28 minuti* on 28/2/2011;¹⁶ and a recording of the original voice by Rossella in the radio broadcast *Un giorno da pecora* on 5/6/2013.

Table 4 shows the formant frequencies of each sub-corpus.

SUB-CORPORA VOICES	WORDS	VOWELS	F1	F2
Giusti-original	<i>ri<u>d</u>ere</i> ('to laugh')	i	169	2094
	<i>se<u>m</u>pre</i> ('always')	e	630	1507
	<i>fa</i> ('he does')	a	640	1388
	<i>lu<u>i</u></i> ('he')	u	343	640
	<i>co<u>n</u>osce</i> ('he knows')	o	498	1234
Rossella target	<i>pre<u>f</u>erita</i> ('preferred')	i	362	2239
	<i>se<u>r</u>a</i> ('evening')	e	404	1893
	<i>oc<u>cu</u>pava</i> ('he was occupying')	a	716	1450
	<i>ba<u>t</u>tuta</i> ('gag')	u	466	714
	<i>do<u>n</u>ne</i> ('women')	o	637	975
Giusti imitating Rossella	<i>ri<u>c</u>co</i> ('rich')	i	251	2312
	<i>de<u>s</u>erto</i> ('desert')	e	605	1768
	<i>pa<u>s</u>sa</i> ('it passes')	a	837	1478
	<i>gu<u>s</u>to</i> ('taste')	u	351	954
	<i>do<u>v</u>e</i> ('where')	o	688	1222

Table 4. Formant frequencies of each Giusti-Rossella sub-corpus. Stressed vowels are underscored.

Below, figures 16-20 show the formant charts and their T, which refer to each Giusti-Rossella sub-corpus.

Figure 16 shows the corresponding formant charts of the voices of Giusti imitating Rossella, of Rossella-target, and of Giusti-original.

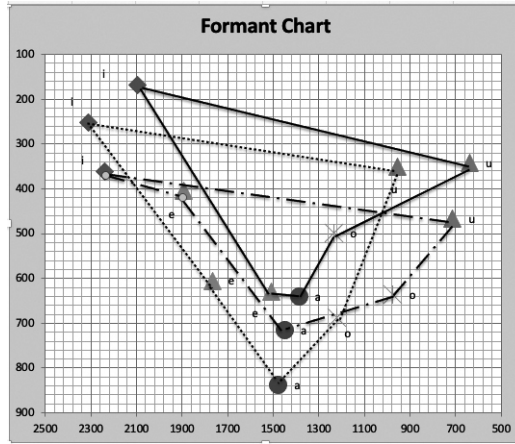


Figure 16. Formant chart of Giusti-Rossella: the dotted line polygon refers to Giusti imitating Rossella, the dash-dot line one refers to Rossella target, and the solid line one refers to Giusti-original.

Figure 17 shows the T from the polygon of Giusti-original to the one of Giusti imitating Rossella. The $SI_1 = 4.765$. But in this case also we must adopt our rejection statement: the polygon of Giusti imitating Rossella (on the right side of figure 17) is patently different from the polygon of Giusti imitating Rossella (in the dotted line of figure 16). Thus, they are not considered as belonging to the same class.

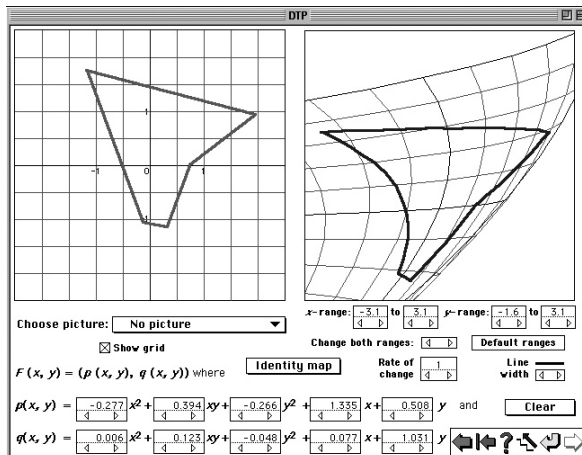


Figure 17. T from the polygon of Giusti-original (left) to the one of Giusti imitating Rossella (right).

Figure 18 shows the T from the polygon of Rossella target to the one of Giusti imitating Rossella. The $SI_2 = 2.728$. But in this case we must adopt our rejection statement: if the polygon T(A) and the polygon B are patently different, then they are not considered as belonging to the same class. In fact, the polygon of Giusti imitating Rossella on the right side of figure 18 is patently different from the polygon of Giusti imitating Rossella in the dotted line of figure 16.

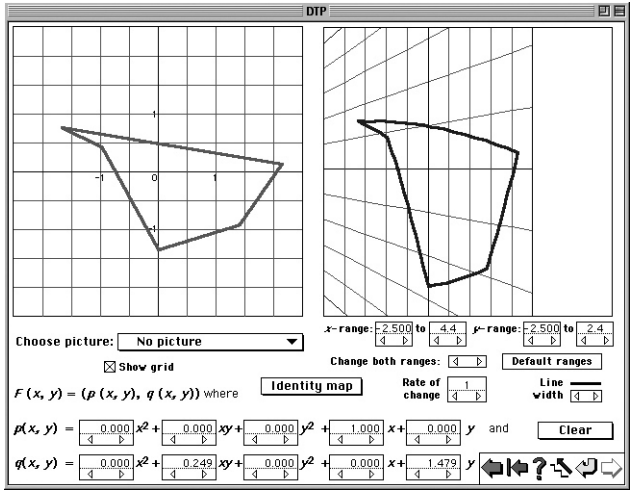


Figure 18. T from the polygon of Rossella target (left) to the one of Giusti imitating Rossella (right).

Figure 19 shows the T from the polygon of Giusti-original to the one of Rossella target. The $SI_3 = 11.395$. But in this case also we must adopt our rejection statement: if the polygon T(A) and the polygon B are patently different, then they are not considered as belonging to the same class. In fact, the polygon of Rossella target (on the right side of figure 19) is patently different from the polygon of Rossella target (in the dash-dot line of figure 16).

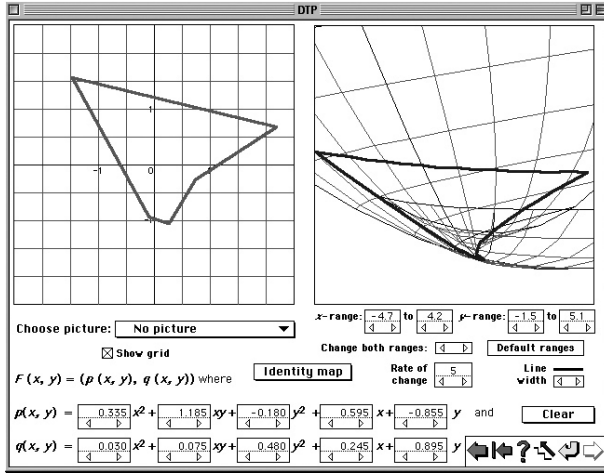


Figure 19. T from the polygon of Giusti-original (left) to the one of Rossella target (right).

Figure 20 shows the inverse-T from the polygon of Rossella target to the one of Giusti-original. The $SI_4 = 2.146$. But in this case also we must adopt our rejection statement: the polygon of Giusti-original (on the right side of figure 20) is patently different from the polygon of Giusti-original (in the solid line of figure 16). Thus, they are not considered as belonging to the same class.

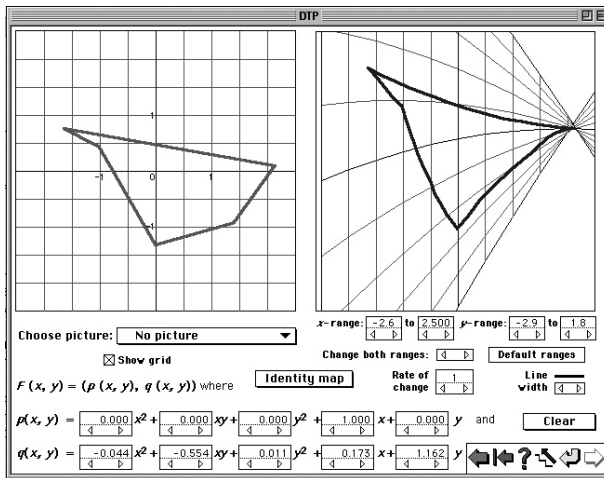


Figure 20. Inverse-T from the polygon of Rossella target (left) to the one of Giusti-original (right).

In the case of Giusti-Rossella the question about similarity is not appropriate. No T is able to pass the rejection statement, and thus their SI is negligible and not significant. Our representation predicts that the voices of the two speakers are not compatible, despite public appreciation.

6. Perception experiment

In order to validate these predictions, we carried out a perception test with a group of listeners.

The speech materials for the experiment are taken from the audio corpus previously described. In particular, the audio is formed by 40 seconds of the recordings of the original voice of each imitated target (Califano, Sordi, Rossella, Lotito) and 40 seconds of the voice of each impersonator imitating the target (Max Tortora, Max Giusti).¹⁷

These recordings were used as stimuli in the perception test presented to 32 listeners: 21 university students from Viterbo (near Rome) and 11 students from adjacent villages. All were native Italian speakers and linguistics students. The mean age was 20.4 (Standard Deviation = 1.1); 12 were male and 20 female. They were not used to listening and judging different kinds of voices but knew they were going to listen to a professional impersonator imitating a given target. As for their familiarity with the targets, as noted earlier, Califano and Sordi are very well known, while Lotito and particularly Rossella are little known.

First, the students listened to the original voice of each imitated target, then to the voices of his impersonator imitating the target. Between every two recordings there was a 5-second pause. The listening test took place in a university classroom. Students listened to the signals by means of a loudspeaker, all together, at the same time, and in the same room; they could listen just once to the signals, in random order.

Immediately after listening to the recordings, the listeners were asked to judge each impersonator on a scale from 1 (similar to the voice of the target) to 10 (not similar to the voice of the target). As a consequence, the closer the total of judgments is to 1, the more similar the voice of the impersonator to its target. They had to submit their answers by filling out a written form.

7. Results of the perception experiment

The results of the perception experiment are given in table 5.

IMPERSONATOR	TARGET	TOTAL POINTS	MEAN	IR
Max Tortora	Franco Califano	33	1.57	2.138
Max Tortora	Alberto Sordi	107	5.09	3.389
Max Giusti	Claudio Lotito	170	8.09	N.A.
Max Giusti	Carlo Rossella	192	9.14	N.A.

Table 5. Results of the perception experiment on a scale from 1 (similar to the voice of the target) to 10 (not similar to the voice of the target). IR value refers to the corresponding topological predictions.

Table 5 shows the correlation between the mean perception rating and the IR of each imitation. The correlation does not apply (N.A.) to both Giusti’s imitations because both imitations did not pass the rejection statement, as noted in §5.3 and in §5.4. The table illustrates that according to the listeners, the voices of Max Tortora and Alberto Sordi or Franco Califano are fully compatible, that is, both imitations are successful, whereas the voices of Max Giusti and Claudio Lotito or Carlo Rossella are not compatible, that is, both imitations are unsuccessful. This is not only because the listeners do not appreciate them, but also because their formant charts did not pass the application of the rejection statement, and this failure shows that there is no possible topological transformation between the formant charts of their voices. More specifically, according to both listeners and IR values, the imitation of Franco Califano by Max Tortora is better than that of Alberto Sordi by the same Max Tortora.

The results of this perception experiment validate the topological predictions. In order to quantify the reliability in perception test, we applied the Cronbach’s Alpha (commonly known as an inter-rater agreement). The Cronbach’s Alpha test gave a global value of 0.997 (total variance = 3825.25). The high value of Cronbach’s Alpha indicates a high reliability level of the perception judgments.

8. Conclusions

The topic of this study is a comparison of two professional voice mimickers (Tortora and Giusti) to determine how successfully they approach the formant values in a few target voices. The number of cases is very limited: two professional Italian impersonators each mimicking two well-known Italian performers/public figures. The study is a ‘method study’ since its primary purpose is to assess if a given method (in De Dominicis 2020) can be used to quantify the degree of appreciation of some imitations.

The method uses formant charts as input forms and applies a topological transformation to them; the formant charts represent the vowel spaces of imitators, imitation, and targets. Based on these transformations, some indices (SI, SI_M, SI_{OV}, and IR) between the compared vowel spaces are calculated.

These similarity indices have been compared with human similarity perception. To this end, a perception test is included in the study. To illustrate the correlation between similarity indices and perception, the IR indices are compared to the estimates based on human perception.

It is worth mentioning that the analyzed impersonations were made for entertainment purposes,¹⁸ and that the inter-rater agreement (Cronbach's Alpha) in the perception experiment is very high. As noted above, this result validates the topological analysis method as a predictor of human speaker similarity perception. Nevertheless, this is a pilot study, and thus by definition further supplementary analyses are required in order to verify this approach.

Abbreviations

DTP: Mac topological application.

IR: Imitation Rank.

N.A.: the relation does not apply.

SI: Similarity Index.

SI_M: mean SI between Mimicking voice and original voices.

SI_{OV}: mean SI between the Original Voices of the impersonator and his target.

Notes

¹ Among them Suzuki (1968) and Zetterholm (2001, 2002a, 2006).

² The paper studied spectrograms of utterances produced by seven speakers and was recorded over periods of up to 29 years.

³ The results by Eriksson & Wretling (1997) and by Wretling & Eriksson (1998) were confirmed by Zetterholm (2002b).

⁴ Majewski (2007), Zetterholm *et al.* (2004), and Zetterholm (2006) came to the same conclusion that F0 is more imitable.

⁵ An example of a formant chart is given in De Dominicis (2020: 64, Figure 2).

⁶ Plotting the values of F1 and F2 in Bark (rather than in Hz) homotopically 'translates' but does not change the overall shapes of both polygons under comparison. Thus, the transfer functions (T) remain the same, both in Hz and in Bark, because in T both the starting functors (or starting polygons) and the final functors (or final polygons) have been homotopically modified.

⁷ A Java Script version of this stack is now available at <mathshistory.st-andrews.ac.uk/Darcy/transformation>.

⁸ As explained in De Dominicis (2020: 68), "this potentially imperfect coincidence between A and A', A'' and A, B and B', B'' and B depends on the operating principle of DTP: since it produces continuous deformations of space, most results actually

derive from the operator's choices, manual skill and manipulations". Thus when the result of T is $B \neq B'$, and $A \neq A'$, the polygons are called 'patently different' and any further measurement of the SI is stopped.

⁹ Here two voices are called 'compatible' if their formant charts are topologically equivalent, i.e. if they belong to the same topological class (cf. De Dominicis 2020: 73). A possible source of 'incompatibility' between voices stems from the application of the 'rejection statement' to the polygons resulting from the transformations in DTP: if the transformation T of a given formant chart A into another B does not give the figure B but something very different (say B'), then A and B are said to be 'patently different' and the 'rejection statement' is applied, that is, any further measurement of SI is stopped (cf. the definitions in De Dominicis 2020: 68-69).

¹⁰ 'Stressed vowels' are vowels with a primary lexical stress. The presence, location, and force (primary) of Italian lexical stress stem from the lexicon. For the restriction of the analysis to stressed vowels, see De Dominicis (2020: 75-76).

¹¹ Of course, in a radio broadcasting recording, the impersonator sometimes uses the same words as his target, sometimes different ones.

¹² In 2005 Tortora was 42 years old and Califano was 67.

¹³ Here and in the following tables, the formant frequencies are considered a simple tool enabling us to plot the formant charts. Thus, the tables include no specific comments on these data.

¹⁴ In 2004 Tortora was 41 years old.

¹⁵ In 2015 Giusti was 58 years old.

¹⁶ In 2011 Giusti was 54 years old.

¹⁷ As for the speakers' dialects, it is worth mentioning that Tortora, Giusti, Califano, Sordi, and Lotito are all from Rome; Rossella is from Pavia (in northern Italy).

¹⁸ This means that the impersonators were not instructed to learn to mimic the targets as closely as possible in advance.

Bibliographical References

- Bessler, Paul 1991. La caricature de de Gaulle par Tissot: Étude phonostylistique. *Information/Communication* 12. 19-32.
- De Dominicis, Amedeo 2020. (Dis)Similarities between formant charts as global topological objects. *Italian Journal of Linguistics* 32,2. 59-98. <DOI: 10.26346/1120-2726-159>.
- Endres, Werner; Bambach, W. & Flösser, Gaby 1971. Voice spectrograms as a function of age, voice disguise, and voice imitation. *Journal of the Acoustical Society of America* 49. 1842-1848.
- Eriksson, Anders & Wretling, Pär 1997. How flexible is the human voice? A case study of mimicry. *Proceedings of the Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997*, Rhodes, Greece. 1043-1046.
- Fant, Gunnar 1960. *Acoustic Theory of Speech Production*. The Hague: Mouton.
- Fant, Gunnar 1966. A note on vocal tract size factors and non-uniform f-pattern scalings. *Speech Transmission Laboratory Quarterly Progress and Status Report*, vol. 1. 22-30.
- Farrús, Cabeceran Mireia; Wagner, Michael; Erro, Eslava Daniel & Hernando, Francisco Javier 2010. Automatic speaker recognition as a measurement of voice imitation and conversion. *International Journal of Speech, Language and the Law* 1. 119-142.

- Hautamäki, Rosa González; Kinnunen, Tomi; Hautamäki, Ville & Laukkanen, Anne-Maria 2015. Automatic versus human speaker verification: The case of voice mimicry. *Speech Communication* 72. 13-31.
- Kitamura, Tatsuya 2008. Acoustic analysis of imitated voice produced by a professional impersonator. *Proceedings of Interspeech 2008, Annual conference of the International Speech Communication Association*. Brisbane (Australia): ISCA. 813-816.
- Majewski, Wojciech 2007. Speaking fundamental frequency of original speakers and their imitators. *Archives of Acoustics* 32,1. 17-23.
- Mejvaldová, Jana 2004. Caractéristiques temporelles de la parole imitée. In Bel, B.; Marlien, I. (eds.), *Actes des XXIVèmes Journées d'études sur la parole (JEP)*. Fès (Maroc). 369-372.
- Perrot, Patrick; Aversano, Guido & Chollet, Gérard 2007. Voice disguise and automatic detection: Review and perspectives. In Stylianou, Yannis *et al.* (eds.), *Progress in Nonlinear Speech Processing*. Berlin: Springer. 101-117.
- Suzuki, Masahisa 1968. Spectrograms: Speaker identification from the viewpoint of voice imitation. *Gengo-seikatsu* 207. 37-41.
- Thompson D'Arcy, Wentworth 1917. *On growth and form*. Cambridge: Cambridge University Press.
- Wretling, Pär; Eriksson, Anders 1998. Is articulatory timing speaker specific? Evidence from imitated voices. In Branderud, Peter; Traunmüller, Hartmut (eds.), *Proceedings of FONETIK 98*. Stockholm: Dept. of Linguistics, Stockholm University. 48-52.
- Zetterholm, Elisabeth 2001. Impersonation: Reproduction of speech. *Working Papers*, Dept. of Linguistics, Lund University, 49. 176-179.
- Zetterholm, Elisabeth 2002a. A comparative survey of phonetic features of two impersonators. *Proceedings of Fonetik, TMH-QPSR* 44,1. 129-132.
- Zetterholm, Elisabeth 2002b. Intonation pattern and duration differences in imitated speech. In Bel, Bernard; Marlien, Isabelle (eds.), *Proceedings of Speech Prosody 2002*. Aix-en-Provence: ISCA. 731-734.
- Zetterholm, Elisabeth 2006. Same speaker – different voices. A study of one impersonator and some of his different imitations. In Warren, Paul; Watson, Catherine I. (eds.), *Proceedings of the 11th Australian International Conference on Speech Science & Technology*. University of Auckland, New Zealand: Australian Speech Science & Technology Association Inc. 70-75.
- Zetterholm, Elisabeth 2007. Detection of speaker characteristics using voice imitation. In Müller, Christian (ed.), *Speaker Classification II. (Lecture Notes in Computer Science)*. Berlin / Heidelberg: Springer. 192-205.
- Zetterholm, Elisabeth; Blomberg, Mats & Elenius, Daniel 2004. A comparison between human perception and a speaker verification system score of a voice imitation. In Cassidy, Steve; Cox, Felicity; Mannell, Robert & Palethorpe, Sallyanne (eds.), *Proceedings of the 10th Australian International Conference on Speech Science & Technology*. Macquarie University, Sydney: Australian Speech Science & Technology Association Inc. 393-397.