UNIVERSITÀ DEGLI STUDI DI PAVIA

DIPARTIMENTO DI STUDI UMANISTICI

CORSO DI LAUREA MAGISTRALE IN LINGUISTICA
TEORICA, APPLICATA E DELLE LINGUE MODERNE

Percezione delle vocali tuderti: uno studio psicoacustico con sintesi formantica

RELATRICE

Prof.ssa Chiara Zanchi

CORRELATORE

Dr. Andrew Murphy

Trinity College Dublin

Tesi di Laurea Magistrale di

Bo Matthew Salvadalena

Matricola n. 548677

Anno accademico 2025/2026

ACKNOWLEDGEMENTS

I am deeply indebted to my two supervisors, Prof.ssa Chiara Zanchi at the Università di Pavia and Prof. Andrew Murphy at Trinity College Dublin, without whom this thesis would not have been possible.

I would like to express my deepest gratitude to Prof.ssa Zanchi for her invaluable guidance, insight, and support throughout the writing of this thesis. In her role as President of the *corso di studi*, she has also been a constant and indispensable point of reference for the practical and bureaucratic aspects of my degree program and of the exchange year at Trinity College Dublin.

Words cannot express my appreciation for Prof. Murphy, whose lectures, ideas, and generosity with his time and expertise were fundamental in shaping the direction of this research. His guidance in designing the experiment and his suggestions regarding the bibliography proved essential at every stage. I am also especially grateful for his efforts in extending my stay at Trinity College Dublin for the final months of this project, time that proved precious for completing and refining this work.

I would like to extend my sincere thanks to Prof.ssa Cotta Ramusino for her support in coordinating the exchange program that made my year in Dublin possible, and to Prof. Marchetti of Trinity College Dublin for his technical guidance and inspiration. I am also grateful to Prof. Gianazza, Prof.ssa Varvara, Prof. Rastelli, and Prof.ssa Zanchi for their willingness to certify at the Università di Pavia the examinations I completed at Trinity College Dublin. Special thanks are also due to Prof.ssa Pavesi for her continued academic guidance throughout my studies.

I would like to acknowledge all the participants who generously contributed their time to the experiment and data collection phase of this research. Their cooperation was essential to this project.

Finally, I would like to thank my family and friends, whose unwavering support and encouragement sustained me throughout this journey.

ABSTRACT

This thesis investigates the tonic vowel system of the Todi variety of Italian from both an engineering and a psycholinguistic perspective. Todi, a hill town in south-western Umbria, belongs to the central Italian dialect zone. Measured data collected for this study show systematic acoustic deviations from Standard Italian. A uniform elevation of the first formant (F_1) is observed across all seven tonic vowels, alongside a front–back asymmetry in the second formant (F_2) that expands the peripherality of the vowel space.

To investigate whether these acoustic differences correspond to differences in listeners' perceptual vowel prototypes, synthetic stimuli are generated. A formant synthesizer is developed with an alias-free Liljencrants–Fant glottal source to produce five-step continua interpolated between the Todi and Standard Italian vowel targets for the seven Italian tonic vowels in Bark space.

Three synthesis conditions (natural, robotic, rough) were evaluated in a MUSHRA naturalness task. Results confirmed that the rough condition was rated poorly when compared with the other two. A forced-choice identification task then confirmed categorization of all seven vowels with an overall above-chance accuracy of 71.8%. A prototype selection task found that listeners from Todi did not select continuum steps closer to the values measured for their dialect. On the contrary, non-Todi, non-Tuscan listeners selected more Todi-like positions. This pattern was most pronounced for the front vowels. These results suggest that speakers from Todi do not have separate mental targets for their vowel inventory than Standard Italian speakers, or that the different acoustic behavior of the Todi speakers is unregistered.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

This thesis investigates the perception of tonic vowels in the variety of Italian spoken in Todi, a central Italian *comune* in the province of Perugia, using a psychoacoustic experiment built around synthetic speech stimuli. The study investigates two related questions: first, the evaluation of the performance of a parallel formant synthesizer and its production of vowel stimuli and their sufficiency for valid perceptual judgments; second, whether listeners from Todi locate the perceptual prototypes of their tonic vowels at different positions along formant continua than listeners from other Italian regions, in a way consistent with the systematic acoustic differences between the two varieties.

The methodological approach is based on an important distinction in speech synthesis between parametric and data-driven systems (Taylor, 2009). Parametric, or rule-based, synthesizers generate speech from explicit mathematical descriptions of the vocal tract: formant frequencies, bandwidths, and amplitudes are set directly by the researcher, and the mapping from parameter to acoustic output is fully transparent. Data-driven approaches instead learn to produce speech from large corpora, using statistical or machine learning models. The state of the art has evolved over time, from approaches such as Hidden Markov Models (Taylor, 2009) to modern neural waveform models (Oord et al., 2016), and these days can achieve striking naturalness. However, the internal parameters of these models, whether statistical distributions or neural network weights, do not correspond to interpretable acoustic quantities. The relationship between input and output is effectively a black box, with no accessible representation of the acoustic signal in terms of formant frequencies or vocal-tract properties. A formant synthesizer, by contrast, operates as a transparent, glass-box system: every parameter has a clear acoustic interpretation, and each can be varied independently while all others are held constant. It is this parametric transparency, and not naturalness, that makes formant synthesis the preferred tool

for acoustic phonetic research, where the objective is to establish causal relationships between specific acoustic dimensions and perceptual responses (Raphael, Borden, and Harris, 2011).

The perceptual component of the experiment is grounded in the prototype theory of phonetic categories and the hyperspace effect. Kuhl (1991) established that vowel categories are organized around internal prototypical exemplars that act as perceptual attractors, compressing the perceived distance between tokens near the prototype and expanding it near the category boundary. Johnson, Flemming, and Wright (1993) showed that listeners' phonemic targets are displaced toward hyper-articulated, peripheral positions that exceed the acoustic mean of naturalistic speech. Since the tonic vowel space of Todi Italian exhibits a characteristic elevation of F_1 and an expansion of F_2 for front vowels relative to the Standard Italian reference documented by Ferrero et al. (1978), the perceptual targets of these two speaker populations are of particular research interest.

1.1 Speech synthesis as a research tool

The use of synthetic speech as experimental stimuli has a long history in speech perception research, as it permits a fundamental methodological advantage: acoustic parameters can be manipulated with a precision and independence that is unattainable with natural recordings. In natural speech, formant frequencies, fundamental frequency, voice quality, duration, and co-articulatory transitions co-vary in ways that cannot be decoupled without fundamentally altering the stimulus. Synthesis, in contrast, allows the experimenter to hold all parameters constant except those under direct investigation. This allows the researcher to attribute perceptual effects unambiguously to the manipulated variables or dimensions (Raphael, Borden, and Harris, 2011). This controllability is the primary rationale for the use of synthetic stimuli in the present study. Since the experimental question concerns the perceptual effect of a precisely defined variation between the Todi and Standard Italian measurements of F_1 and F_2, the stimuli must vary along that dimension alone, with voice quality, duration, and higher formants held constant across all continuum steps.

The use of synthesis to investigate vowel perception begins with the work done

by Cooper et al. (1952), who demonstrated with a hand-painted pattern playback synthesizer that listeners could reliably identify vowels from stimuli containing only two spectral bars positioned at the approximate locations of F_1 and F_2. This result established that F_1 and F_2 are jointly sufficient (and individually necessary) for vowel identity. This finding was subsequently confirmed and extended to natural speech by Peterson and Barney (1952) across a large and phonetically diverse sample of American English speakers. These and subsequent studies demonstrate the suitability of the F_1–F_2 plane as a parameter space for vowel continuum construction and provide the empirical basis for the two-formant interpolation design described in Section 2.4.

The principal objection to synthetic stimuli in perception experiments is that listeners may respond to the artificial quality of the signal rather than to its phonetic content, producing results that reflect characteristics of the synthesizer rather than the listener's phonological categories. This concern is not hypothetical: D. H. Klatt and L. C. Klatt (1990) demonstrated that voice quality parameters, including jitter, shimmer, and spectral tilt, exert measurable effects on vowel naturalness ratings and can interact with phonetic category judgments. For this reason, the present implementation replaces the simple impulse train excitation used in early formant synthesizers with the LF (Liljencrants-Fant) glottal flow derivative model (Fant, Liljencrants, and Lin, 1985; Gobl, 2021). The impulse train produces a spectrally flat glottal excitation that, while adequate for intelligibility, yields a characteristically robotic quality that listeners may not accept as a natural vowel. The LF model, in contrast, generates a waveform with physiologically realistic open-phase and return-phase dynamics whose spectrum more closely approximates natural modal phonation. The validity of the synthesis is not just assumed but also empirically tested: the MUSHRA task (Section 3.7) provides a direct measure of the perceived naturalness for each synthesis condition against a natural speech reference. Furthermore, a forced-choice identification task (Section 3.8)) verifies wether the synthetic vowels are categorized consistently with their intended phonetic targets.

The vocal tract filter is implemented as a formant synthesizer, in which the vocal

tract transfer function is approximated by a bank of second-order digital resonators, each tuned to a formant frequency and bandwidth, as proposed by D. H. Klatt (1980). Two standard topologies exist. In the cascade synthesizer, resonators are connected in series so that their transfer functions multiply: this arrangement has the advantage of self-calibrating formant amplitudes through the natural interaction of adjacent resonances, without requiring explicit gain adjustment for each formant. However, it is less well suited to perception experiments, because formant amplitudes are not independently controllable: any change to F_1 or its bandwidth propagates through the amplitude structure of all higher formants (D. H. Klatt, 1980). In contrast, the parallel synthesizer connects resonators so that their outputs sum, and assigns an independently settable gain parameter to each branch, allowing formant amplitudes to be matched precisely to a reference spectrum. Holmes (1983) argues on these grounds that the parallel architecture is the preferable configuration for speech perception research, where the correspondence between synthesis controls and measurable acoustic properties must be exact. For the present study, which requires reproducible formant amplitude calibration across a five-step continuum and across three qualitatively distinct excitation conditions, the parallel architecture is adopted.

For this experiment a parallel formant synthesizer is developed using an alias-free LF glottal source to provide a controlled and parametrically transparent system for generating synthetic vowel stimuli. Its principal advantages over natural recordings (parameter independence, reproducibility, and the ability to interpolate smoothly between acoustic endpoints) are exploited by the continuum construction procedure described in Section 2.4. The complete synthesis pipeline is presented in Chapter 2 and covers the source-filter model (Section 2.1), the cascade and parallel filter implementations with gain calibration (Section 2.2), the LF source model and its jitter, shimmer, and tremor perturbation parameters (Section 2.3), and the Bark-space interpolation procedure (Section 2.4).

1.2 Research questions

This thesis considers two related research questions, one from an engineering perspective and the other linguistic. The engineering question concerns the perceptual adequacy of the synthesis pipeline. The synthesizer must produce stimuli of sufficient naturalness such that listeners can engage with them as phonetic objects rather than as artifacts of the synthesis process. The linguistic question, as the primary scientific investigation of this work, concerns the representation of vowel prototypes in Standard Italian listeners compared with those from Todi. For this reason, the linguistic research is interpretable only after the stimuli engineering has been shown to be satisfactory.

The engineering question is more specifically: does the developed speech synthesis system (parallel formant synthesizer; alias-free LF glottal source; jitter, shimmer, and tremor) produce tonic vowel stimuli of sufficient naturalness to support valid perceptual judgments, or does it introduce systematic artifacts that could bias listeners' responses? Three synthesis conditions (robotic, natural, and rough) are evaluated against a natural speech reference in a MUSHRA evaluation (International Telecommunication Union, 2015). The hypothesis is that the natural condition will exceed the robotic and rough conditions in rated naturalness and approach the reference.

The linguistic question investigates if listeners from Todi locate their perceptual prototype (the location in the F_1-F_2 space they consider most representative of each tonic vowel in their own dialect) at a different position from Standard Italian listeners along a continuum between the Todi measurements and those of Ferrero et al. (1978). Kuhl (1991) established that vowel categories are organized around internal prototypes that act as perceptual attractors ("magnets"), compressing the perceptual space near the category center and expanding it at the periphery. Johnson, Flemming, and Wright (1993) offer compelling evidence for a phenomenon that they deem the 'hyperspace effect', which indicates that listeners' internal prototypes are displaced towards hyper-articulated peripheral targets that extend past measured values collected from laboratory speech samples. These findings predict that Todi listeners would locate their prototypes closer to the Todi targets, while Standard

Italian listeners would favor positions closer to the Standard Italian targets. The magnitude of this divergence is expected to vary across vowels in proportion to the perceptual distances between the two varieties' targets.

For the prototype selection results (Section 3.9) to be interpretable as evidence for genuine perceptual representations, the synthesis must be sufficiently natural to elicit phonetic responses. The MUSHRA evaluation (Section 3.7) and forced-choice identification task (Section 3.8) verify these preconditions before the main task is interpreted. The experimental design and results are presented in Chapter 3 and Chapter 4.

1.3 Thesis overview

The remainder of this thesis is structured to address the two research questions identified in Section 1.2; from the engineering approach to speech processing that makes the experiment possible, through the experimental design, to the empirical results and their phonological and cognitive interpretation.

Chapter 2 presents the synthesis system: source-filter theory, the parallel and cascade vocal tract filter implementations with gain calibration, the LF glottal source model and its jitter, shimmer, and tremor parameters, and the Bark-space continuum synthesis procedure.

Chapter 3 presents the experimental methodology. The phonological background on Standard Italian and the Todi variety is provided in Sections 3.1 and 3.2, followed by the Todi speaker data collection and acoustic analysis (Sections 3.3 and 3.4), the theoretical framework for the perception experiment (Section 3.5), the continuum design (Section 3.6), and the three experimental tasks: MUSHRA quality evaluation, forced-choice identification, and prototype selection (Sections 3.7 to 3.9).

Chapter 4 reports the results of all three tasks and discusses their implications.

Chapter 5 draws overall conclusions, identifies the principal limitations, and proposes directions for future work.

Chapter 2

IMPLEMENTATION

2.1 Source-filter theory

The acoustic theory of speech production, formulated by Fant (1960), proposes that the speech signal can be decomposed into three separable components: a sound-generating source at the larynx, a resonant vocal tract filter that shapes the source spectrum, and a radiation characteristic that models the acoustic transformation occurring as sound radiates from the lips. In the frequency domain, these three contributions combine multiplicatively, so that the output speech spectrum $S(f)$ is expressed as:

$$S(f) = U_g(f) \cdot T(f) \cdot R(f) \tag{2.1}$$

where $U_g(f)$ is the glottal volume velocity spectrum, $T(f)$ is the vocal tract transfer function, and $R(f)$ is the radiation characteristic. The mathematical relationship between source and filter draws on signals and systems theory for the convenient practical consequences of separating speech production into components. This separation has profound consequences for synthesis design, as will be shown.

The most important theoretical result of the source-filter model is the assumed independence of the source and the filter. The laryngeal source (both during the quasi-periodic vibration of the vocal folds during voicing and the production of turbulent noise generated with a constriction during frication) is generated by mechanisms that are, approximately, unaffected by the configuration of the overlying vocal tract. Conversely, the supra-laryngeal vocal tract resonates at frequencies determined by its geometry, independently of the mode of laryngeal activity. Raphael, Borden, and Harris (2011) illustrate this independence through the following simple example: speakers can smoothly glide from one pitch to another while sustaining a fixed vowel and simultaneously alter the source (fundamental frequency) while holding the filter (formant structure) constant. The two systems therefore behave

like a linear time-invariant (LTI) system. This in turn allows the formulation of the output speech spectrum as a series of products in the frequency-domain as in Eq. (2.1).

Independence allows for three main advantages for digital synthesis. First, the vocal tract can be modeled as a linear time-invariant filter driven by an independent excitation signal. This system design reduces a complex biological process into a standard digital signal processing and filtering problem. Second, source parameters (e.g. fundamental frequency, voice quality, etc.) and filter parameters (e.g. formant frequencies, bandwidths, and amplitudes) can be manipulated independently without mutual interference. This provides much more precise control over experiments than a monolithic system. Third, a single digital filter configuration can be powered by qualitatively different source signals without restructuring the filter itself. For instance, a periodic pulse train for voicing or broadband noise for frication. The vocal tract filter $T(f)$ is characterized acoustically by a set of resonant peaks called formants. Formants are the natural resonant frequencies of the vocal tract, which are created when sound waves bounce back and forth inside the mouth and throat to amplify specific frequencies. For simplicity of calculation, the vocal tract is often modeled as a concatenation of uniform acoustic tubes. As a result, its transfer function contains only poles. Its denominator polynomial is therefore a product of complex conjugate pole pairs, each corresponding to a single formant (D. H. Klatt, 1980). Digitally, each conjugate pair is realized as a second-order infinite impulse response (IIR) resonator. This serves as the basic building block of formant synthesis, described in detail in Section 2.2. While the human vocal tract possesses resonances extending well above the range of intelligibility, Holmes (1983) notes that five resonators adequately model the spectral envelope up to 5 kHz, the perceptually relevant range for vowel quality in an adult male vocal tract. The synthesizer developed for this thesis thus limits its output to this range of frequencies.

The radiation characteristic $R(f)$ describes the physical behavior of sound as it leaves the speaker's lips and travels through air. Because higher frequencies radiate from the head more efficiently, this process behaves similarly to a high-pass filter.

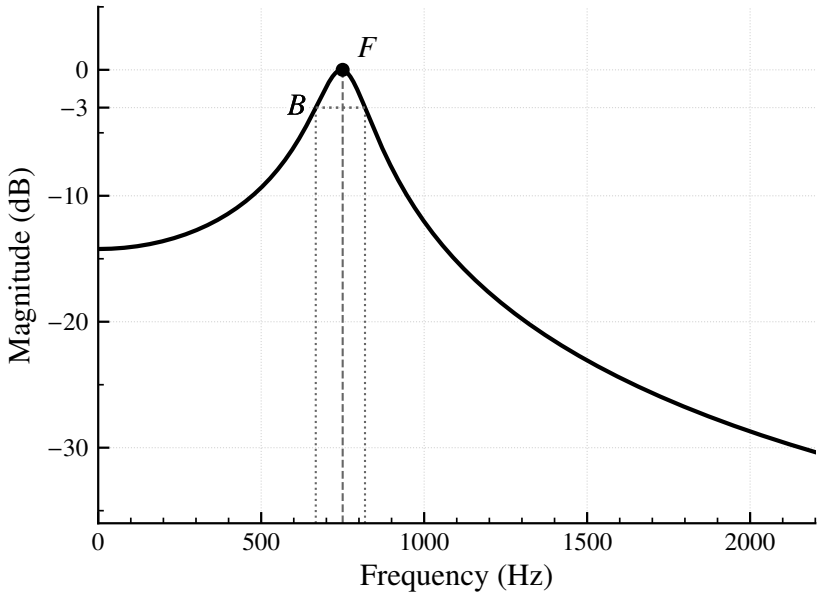

Figure 2.1: Frequency response of a second-order resonator, showing formant peak F and bandwidth B.

Practically, this implies an increase in magnitude of +6 decibels per octave. However, in the present implementation, this term is not modeled as an explicit filter, as it is already accounted for in the differentiated glottal flow waveform generated by the LF model. Given that the vocal tract is treated as a LTI system, the order in which its components are cascaded does not affect the overall system response (Oppenheim, Willsky, and Nawab, 1996).

It should be noted that the source-filter independence assumed by Eq. (2.1) is an assumption that models the phonation process rather than reproducing it exactly. In reality, the assumption of linearity is "fulfilled in essentials only" (Fant, 1995). The phonation process is only roughly linear, as pressure variations within the vocal tract can influence the airflow from the vocal folds. Furthermore, the filtering properties of the vocal tract itself change rapidly even during a single cycle of vocal fold vibration. When synthesizing vowel sounds, these effects have some importance and can produce audible differences if ignored entirely. In any case, these minor

effects fall outside the purpose and scope of the present work and can therefore be neglected.

To summarize, source-filter theory provides a mathematically elegant and transparent model for vowel production that closely approximates the acoustic properties of human phonation. In the following sections, the vocal tract filter in cascade and parallel implementations is examined (Section 2.2); the glottal source model and its perturbation parameters are described (Section 2.3); and the procedure for generating perceptually graded continua from these components is detailed (Section 2.4).

2.2 The vocal tract filter

The vocal tract filter is the component of the synthesis system directly responsible for shaping the formant structure and spectral envelope of the synthetic vowel stimuli. Two complementary filter architectures are employed in sequence. A five-formant cascade synthesizer model is computed first: the cascade topology incorporates an implicit higher-pole correction that produces theoretically accurate formant amplitudes without requiring explicit gain specification for each resonance (Holmes, 1983). However, because the perceptual continua in this study are defined exclusively in the F_1-F_2 plane, the spectral contribution of the upper formants must be withheld from listeners. Truncating a cascade filter directly would disturb the amplitudes of the lower formants, since their amplitudes are dependent on higher resonances. To avoid this, the cascade output is used not as the final stimulus but as a calibration tool. The formant amplitudes it produces are extracted and used to set the per-formant gain parameters of a three-formant parallel synthesizer model. Because the parallel architecture affords independent control over each formant amplitude, it can reproduce the F_1–F_3 spectral envelope of the cascade exactly while excluding all higher-formant information from the output. F_3 is retained in the stimulus alongside F_1 and F_2 due to its importance for accuracy in vowel identification tasks (Hillenbrand et al., 1995). Without F_3, the second formant would have to be artificially raised to a position very close to the third formant in order to generate convincing front vowels (Raphael, Borden, and Harris, 2011).

The transfer function class

The transfer functions of the two synthesizers are implemented in the frequency domain through a combination of second-order digital resonators. Because Python lacks a dedicated package for manipulating pole-zero transfer functions, a custom object-oriented `TF` class was developed. The class stores the numerator and denominator polynomial coefficients of a rational transfer function $H(z) = B(z)/A(z)$. This class serves as an abstract representation of the underlying mathematics that enables direct arithmetic operations on transfer functions in accordance with classical signal-processing theory. Multiplication of two transfer functions (cascade connection) is carried out through convolution of their respective numerator and denominator coefficient arrays. Addition (parallel connection) follows the algebraic procedure for summing fractions: the numerators are cross-multiplied with the opposite denominators, the resulting numerator polynomials are added, and the common denominator is obtained by convolution of the individual denominators. Operator overloading simplifies the syntax so that transfer functions combine using familiar Python notation. Full code listings of the `TF` class and all related helper functions are provided in Appendix A.

Digital resonator design

Individual second-order formant resonators are constructed with a dedicated function `resonator`. Given a formant frequency F_n, bandwidth B_n, and sampling period $T_s = 1/f_s$, the pole radius and angle are computed as:

$$r = e^{-\pi B_n T_s} \tag{2.2}$$

$$\theta = 2\pi F_n T_s \tag{2.3}$$

The resulting transfer function takes the standard digital resonator form (D. H. Klatt, 1980):

$$R(z) = \frac{A}{1 - Bz^{-1} - Cz^{-2}} \tag{2.4}$$

where coefficients A, B, C derive directly from the equations given in D. H. Klatt (1980), that is, $B = 2r\cos\theta$, $C = -r^2$, and $A = 1 - B - C$. This produces a resonance peak at frequency F_n with bandwidth B_n.

Cascade system

For vowel sounds, the cascade connection models the vocal tract transfer function by connecting the individual resonators R_n in series. The continuous-time all-pole transfer function of Fant's (1960) acoustic tube model of the vocal tract is:

$$H(s) = \prod_{n=1}^{\infty} \frac{s_n s_n^*}{(s - s_n)(s - s_n^*)} \tag{2.5}$$

While the number of poles is infinite, resonances above 5 kHz carry very little importance to the output signal. In particular, both the power spectral density of the sound source and the sensitivity of human hearing are reduced above this range. In digital systems, the z-domain transformation automatically produces an infinite series of poles. As a result, it is sufficient to explicitly model poles up to the Nyquist frequency for automatic higher pole correction (Holmes, 1983). The resulting discrete-time transfer function of the vocal tract filter is therefore:

$$H(z) = \prod_{n=1}^{N} R_n \tag{2.6}$$

where R_n is the n-th formant resonator. In the present implementation, the function `cascade` returns the overall transfer function of the concatenated resonators as a `TF` object.

Parallel system and gain calibration

The parallel architecture feeds the input signal simultaneously into all resonators and sums their outputs. Because the transfer functions are combined additively rather than multiplicatively, the natural higher-pole correction of the cascade model is absent: formant amplitudes do not self-correct and must be controlled explicitly through individual gain parameters. The overall parallel transfer function is:

$$H(z) = \sum_{n=1}^{N} (-1)^n R_n G_n H_{\text{filter}} \tag{2.7}$$

where, for each formant n, G_n is the gain factor, H_{filter} is a spectral shaping filter, and $(-1)^n$ is a scalar coefficient for alternating polarity. Since the importance of

gain control has already been addressed, the reasoning underlying the choice of alternating polarity and spectral shaping filters will now be discussed.

Two spectral-shaping operations are required to produce clean vowel spectra. A pre-emphasis filter

$$H(z) = 1 - \alpha z^{-1} \tag{2.8}$$

is applied to the F_1 branch and a first-difference differentiator

$$H(z) = 1 - z^{-1} \tag{2.9}$$

is applied to all higher formants. These filters remove the low-frequency skirt energy from the higher resonators that would otherwise bleed into the F_1 region and distort the spectrum. Polarity alternation $(-1)^n$ is applied across all branches. Without it, destructive interference between parallel branches would produce deep spectral notches in the valleys between formant peaks. The alternating sign (+/−) mitigates this effect and yields smoother combined spectra (D. H. Klatt, 1980).

Gains, used to match amplitudes with the cascade reference, are calculated in a two-step process. First, the parallel system is calculated with unity-gain. From this first model, the peak magnitudes at each target formant frequency are extracted. Subsequently, the required gains are then computed as the ratio of the cascade peak amplitudes to the corresponding unity-gain parallel peaks. These gains are then passed to the `parallel` function. This aligns the formant peaks of the parallel model with those of the cascade reference within 0.1 dB.

Despite the calibration overhead, Holmes (1983) argues that the parallel architecture is superior to the cascade for speech perception research. Because the amplitude of each formant is an explicit, independently controllable parameter, the model's controls correspond directly to measurable acoustic properties. Formant frequency, bandwidth, and amplitude can therefore be verified and adjusted against spectral measurements of natural speech with maximum control. Holmes (1983) further demonstrates that parallel synthesizers are able to handle all speech sounds monolithically, whereas cascade synthesizers require separate control parameters for consonants such as fricatives and plosives, which introduces discontinuities at phonetic

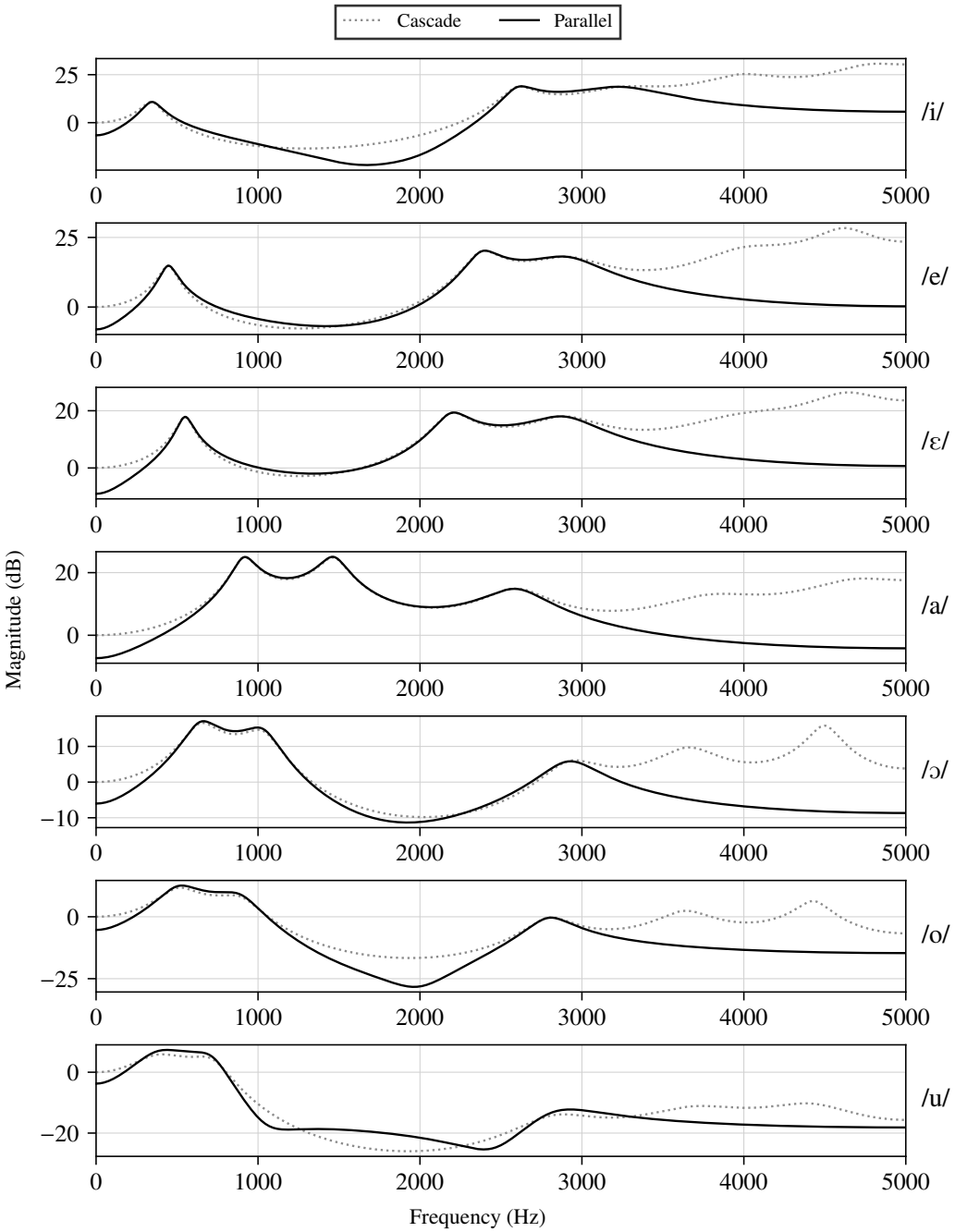

Figure 2.2: Frequency responses of the transfer functions for each of the seven vowels derived from mean formant frequencies and bandwidths measured across eight speakers from Todi. Five resonators (F_1–F_5) are used to generate the cascade filter (dotted). The parallel filter (solid) is then gain-calibrated to match (F_1–F_3).

boundaries. In the context of psycho-acoustic research, this is relevant for the generation of whole words and connected speech. The present experiment, however, only requires the synthesis of steady-state vowels. For that reason, the precise formant amplitude control of the parallel model is more suitable for the constraints of the continuum stimuli.

Comparison of systems

The magnitude responses of the cascade and parallel vocal tract transfer functions show excellent agreement in the location and amplitude of the formant peaks across all seven vowels. This confirms that the computed gain array successfully aligns the parallel model with the natural resonance heights of the cascade reference. Peak deviations remain below 0.1 dB for the target formants.

Table 2.1: Center frequencies (f, Hz) and bandwidths (B, Hz) for F_1–F_3 of each of the seven Todi vowels.

	F_1		F_2		F_3	
Vowel	f	B	f	B	f	B
/i/	344	108	2601	164	3236	489
/e/	446	92	2380	177	2912	407
/ɛ/	549	86	2193	190	2887	395
/a/	912	126	1467	157	2594	348
/ɔ/	643	188	1032	204	2918	322
/o/	505	219	892	243	2794	246
/u/	378	320	717	227	2825	453

The vowel /o/ exhibits the most pronounced divergence between the two synthesis models. While the cascade response rolls off gradually above F_2 (892 Hz), the parallel develops a sharp trough below −25 dB near 2000 Hz before recovering at F_3 (2794 Hz) as a result of the sign alternation between adjacent parallel branches. The vowel /u/, overall, presents the most heavily attenuated spectrum: the wide F_1 bandwidth of 320 Hz rounds the first resonance into a shallow peak barely above 0 dB, and the response remains at or below −15 dB across the entire region above 1000 Hz. In /i/, a deep valley spanning roughly 800 to 2000 Hz separates a moderate

F_1 peak of approximately 12 dB at 500 Hz from the F_2–F_3 region, where the two resonances (2601 and 3236 Hz) merge into a broad plateau near 22 dB with only a shallow dip between them rather than two sharply resolved peaks.

Compared to the five-resonator cascade model, the parallel model (calibrated to match only F_1–F_3) exhibits a steeper spectral roll-off above F_3, as expected.

2.3 The glottal source

The vocal tract filter described in the previous section shapes the spectral envelope of the output, but it is only as good as the signal that drives it. In formant synthesis, that driving signal is a model of the glottal airflow pulse, which determines the overall spectral slope and perceived voice quality of the result. A simple periodic impulse train produces technically correct formant structure but sounds unmistakably synthetic; the regularity is inhuman. Constructing a realistic source therefore requires tackling two separate layers of the problem: first, the shape of the individual glottal pulse, and second, the irregularity of the real human voice.

The LF model

The standard model for the glottal source in formant synthesis is the LF model, introduced by Fant, Liljencrants, and Lin (1985). Rather than modeling the glottal airflow directly, the LF model specifies its time derivative. When the time derivative is convolved with the vocal tract transfer function the output speech signal is produced. The LF model is particularly suited to speech synthesis and analysis because its key acoustic events are easy to parameterize.

Each glottal cycle is described by two piecewise segments. During the open phase, when the vocal folds are opening and closing, the flow derivative follows an exponentially growing sinusoid that rises to a negative peak at the main excitation instant T_e. From that peak, the return phase brings the waveform back to zero through an exponential decay governed by the return time constant T_a.

The four original model parameters were subsequently reformulated by Fant (1995) into a single parameter R_d. The R_d value captures the dominant dimension of co-

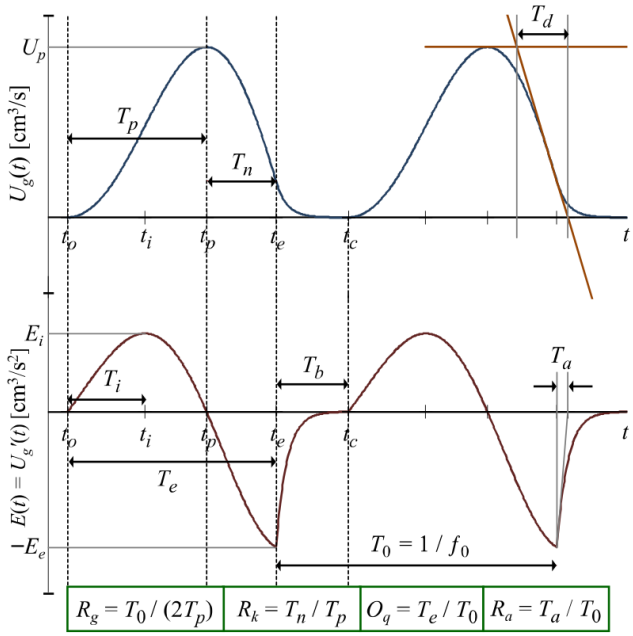

Figure 2.3: The LF-model from Gobl (2017).

variation among the original parameters and ranges from approximately 0.3 (pressed, tense phonation) to 2.7 (breathy, lax phonation) for normal voices. Increasing R_d raises the open quotient and lengthens the return phase. As a consequence, the source spectrum has a gentler roll-off and a stronger fundamental relative to its harmonics. Listeners hear this as increased breathiness. Decreasing R_d has the opposite effect. The R_d parameter is therefore the primary control handle for voice quality in the present synthesis system.

Generating an LF pulse for a target R_d value is not straightforward, because the mapping from parameters to R_d is nonlinear. Gobl (2017) provides an iterative Newton–Raphson algorithm that solves for the underlying parameters while enforcing the required R_d, which ensures convergence across the full permissible range of voice qualities.

Alias-free frequency-domain implementation

Sampling an LF pulse in the time domain introduces aliasing: energy above the Nyquist frequency folds back into the spectrum. Gobl (2021) eliminates this problem by computing the pulse spectrum analytically from its Laplace transform and sampling it directly in the frequency domain. This technique produces an alias-free LF pulse without distortion from aliasing. Aliasing effects are perceptually, having been confirmed by Wang and Gobl (2023), whose listening experiments showed that the distortion is reliably detectable even at 20 kHz sampling rates and becomes more salient as fundamental frequency increases. For that reason, present implementation uses the frequency-domain method of Gobl (2021) throughout.

Jitter, shimmer, and tremor

A perfectly periodic LF source still sounds robotic, as human phonation is never perfectly periodic. This noise is perceptually relevant to a natural sounding voice (D. H. Klatt and L. C. Klatt, 1990). For the purposes of the present study, three distinct perturbation phenomena will be considered.

Jitter is random fluctuations in the fundamental frequency. In normal modal voice, jitter is small: Schoentgen (2003) reports median relative jitter of approximately 0.4% of the cycle length across a corpus of 114 sustained vowel utterances from healthy speakers. At these levels, jitter contributes the slight organic roughness of a natural voice. During phonation, the laryngeal musculature contracts at intervals that are approximately (but never exactly) regular (Schoentgen, 2001; Schoentgen and Aichinger, 2019).

Shimmer is random variation in the peak amplitude of successive glottal pulses. Typical values in normal voice are in the range of 1–3% of mean amplitude (Farrús, Hernando, and Ejarque, 2007). Like jitter, moderate shimmer is an expected feature of healthy phonation; it is measurable even in professionally trained voices. D. H. Klatt and L. C. Klatt (1990) demonstrated through analysis, synthesis, and listening experiments that both jitter and shimmer contribute to perceived naturalness, and that artificially removing them from natural speech makes it sound more robotic.

Tremor is a slow, quasi-sinusoidal modulation of fundamental frequency and amplitude simultaneously. This is caused by vibrations in the laryngeal muscles at frequencies typically between 4 and 6 Hz (Schoentgen, 2003). In normal voices, tremor is present at low amplitudes and produces a subtle wavering in sustained vowels. Listeners associate this wavering with living phonation rather than robotic synthesis. At elevated amplitudes or irregular rates, tremor is pathological, however, at normal levels, it is simply a component of the natural voice pattern.

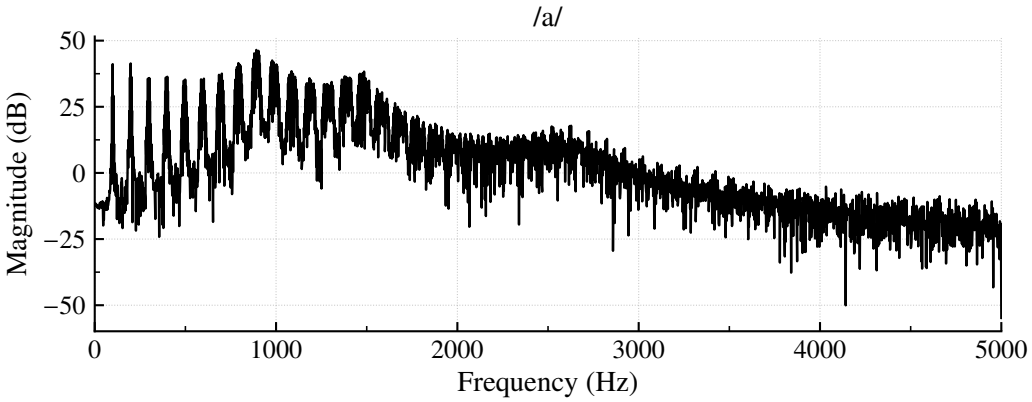

Figure 2.4: Spectrum of the parallel formant synthesizer for the Todi /a/ stimulus.

These three parameters contribute in varying degrees to shaping the glottal source. In the three source conditions used in this experiment, their levels are varied to span a perceptual range from wholly unperturbed to naturally perturbed to excessively rough. The MUSHRA naturalness evaluation in Section 4.1 tests how listeners respond to this parametric range.

2.4 Continuum synthesis

The Bark scale and its perceptual motivation

The perception experiment requires stimuli that form a graded continuum between the two phonetic targets: on one end, the vowels from Todi, and on the other the Standard Italian vowels from Ferrero et al. (1978). A simple implementation would be to linearly interpolate the target formant frequencies in hertz. However, this

is not a satisfactory solution, as the relationship between physical frequency and tonotopic nature of the cochlea is nonlinear. Because of this, the Hz scale does not yield perceptually equal steps. Equal-sized steps in hertz therefore correspond to unequal perceptual distances, particularly at the lower end of the formant frequency range where the tonotopic scale compresses rapidly. For this reason, interpolation is instead performed on the Bark scale, which approximates the tonotopic organization of the human auditory apparatus.

Traunmüller (1990) explains that the Bark scale maps to areas where different frequencies stimulate the inner ear. Since this mapping represents the physical distance between them inside the ear, the Bark scale is well suited for perception experiments. Traunmüller (1990) derives the following approximation for the critical-band rate z in Bark as a function of frequency f in Hz:

$$z = \frac{26.81\, f}{1960 + f} - 0.53 \tag{2.10}$$

The inverse transformation, required to convert synthesized Bark values back to formant frequencies in Hz, is:

$$f = \frac{1960\,(z + 0.53)}{26.81 - z - 0.53} \tag{2.11}$$

These expressions are accurate to within ±0.05 Bark over the range 200 Hz–6.7 kHz, which encompasses all formant frequencies of interest in the present study.

Having established that the Bark scale is more suitable than a scale in hertz, the step size between each of the stimuli on the continuum must now be considered. To that end, the just discriminable difference for formant frequency serves as the lower bound for meaningful step size. Flanagan (1957) established, through psychoacoustic experiments with synthetic vowels, that the discrimination threshold for formant frequency is approximately ±3% of the formant value. For typical adult male formant values this corresponds to absolute precisions of ±20 Hz for F_1, ±50 Hz for F_2, and ±75 Hz for F_3. In the Bark domain, ±3% at a representative F_1 of 400 Hz is equivalent to approximately 0.22 Bark. The five-step continuum described in the following section yields per-step Bark increments ranging from 0.21 Bark for /u/ to

Table 2.2: Euclidean distances and per-step increments in Bark between Todi and Standard Italian tonic vowels on the five-step continua.

	Todi (Hz)		SI (Hz)		Bark	
Vowel	F1	F2	F1	F2	Distance	Step
/i/	344	2601	290	2310	0.957	0.239
/e/	446	2380	350	2050	1.348	0.337
/ɛ/	549	2193	490	1950	0.934	0.234
/a/	912	1467	780	1430	0.897	0.224
/ɔ/	643	1032	550	970	0.835	0.209
/o/	505	892	390	870	1.053	0.263
/u/	378	717	320	800	0.822	0.205

0.34 Bark for /e/, which places each adjacent pair at or above the JND. This design ensures that every step is individually discriminable to listeners. Furthermore, it also limits the total number of stimuli to a reasonable quantity for a single experimental session in order to prevent participant fatigue.

Linear interpolation in Bark F_1–F_2 space

The five continuum steps are generated by dividing the total distance between the Todi and the Standard Italian targets by four. The Todi values are derived from the acoustic measurements of eight Todi speakers described in Chapter 3; the Standard Italian values are taken from Ferrero et al. (1978). The total Bark distance between the targets varies across vowels; Table 2.2 reports the per-vowel distances and the resulting per-step increments. The range spans from 0.822 Bark for /u/, whose Todi and Standard Italian measurements are closest in the F_1–F_2 plane, to 1.348 Bark for /e/, which presents the greatest separation. In all cases the steps exceed the JND for the respective frequencies when calculated according to the parameters given by Flanagan (1957).

Fixed higher formants

Only F_1 and F_2 are varied along the continuum. This restriction follows from the established primacy of the first two formants in determining vowel identity: F_1 and

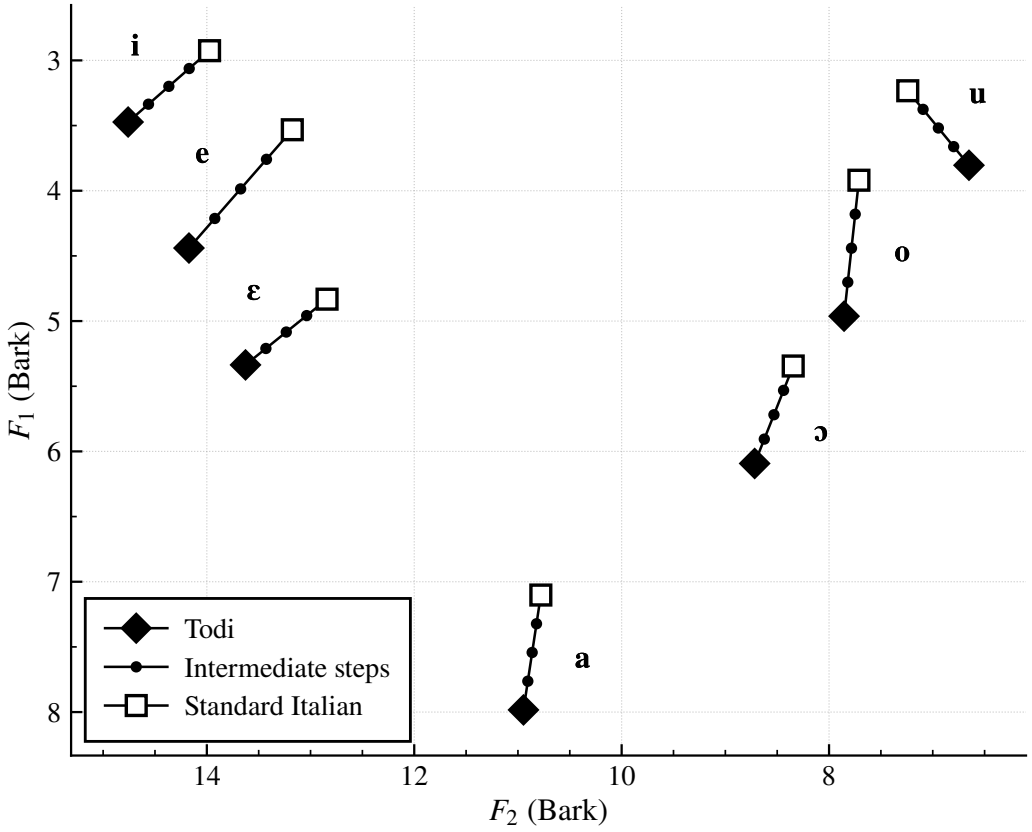

Figure 2.5: Five-step continua between Todi and Standard Italian targets in Bark space.

F_2 account for the principal dimensions of tongue height and tongue advancement, and plots of F_1 versus F_2 reliably separate the vowels of most languages into distinct perceptual clusters (Peterson and Barney, 1952; D. H. Klatt, 1980). It is along these two dimensions that the Todi and Standard Italian vowel systems principally differ; higher formants are consequently fixed to prevent the introduction of a confusing third dimension of variation.

The fixed values assigned to each higher formant across all five steps are as follows. F_3 is held at the Todi-measured value for the respective vowel. This choice maintains acoustic consistency with the reference tokens used in the forced-choice

identification task (Section 3.8) and avoids a systematic spectral shift in the upper vowel region that could interact with listeners' identification responses. F_4 is fixed at the mean of the Todi and Standard Italian measured values. This provides a neutral midpoint that does not pull the overall spectral character of the stimulus toward either endpoint across the continuum. F_5 is fixed at 4500 Hz for all vowels and all steps. At this frequency, vowel-specific variation is perceptually negligible, and the fixed value chosen approximates the mean position of this resonance in adult male speech (Fant, 1960). As Ferrero et al.'s (1978) data does not include bandwidth measurements, formant bandwidths are held constant at the Todi-measured values throughout all five steps. Holding the bandwidths constant ensures that any perceptual differences between steps are attributable solely to the shift in formant center frequency, without confounding contributions from changes in resonance sharpness or overall spectral level. The effect of bandwidth on peak amplitude is, however, not trivial and a limitation of this approach (D. H. Klatt, 1980). In the absence of bandwidth data for the Standard Italian targets, it is nonetheless a reasonable choice, assuming that the bandwidths between the two varieties do not diverge substantially.

Stimuli

The three experiments, MUSHRA, identification, and prototype selection, require different stimuli configured to the methodological constraints of the individual task. First, as mentioned previously, the MUSHRA task requires different stimuli for the three sound-quality or naturalness conditions: natural, robotic, and rough. Across the seven vowels, this leads to $3 \cdot 7 = 21$ total stimuli. This number increases to 24 when the set of natural speech recordings extracted from the Todi speaker data collection used as references are included. For the MUSHRA task the Todi targets are used. Second, for the forced-choice identification task, seven reference vowels are synthesized at the Todi targets under the natural condition. Third, for the prototype selection task, the natural setting is also used for all of the continua. The details of the variation in the F_1–F_2 Bark space are discussed above; to summarize, the five equidistant perceptually discriminable steps are interpolated in Bark space between the Todi and Standard Italian targets while maintaining higher formants

constant.

All of the stimuli are 250 ms long with a 10 kHz sampling rate and stored as 16-bit PCM WAV files. For the MUSHRA task, each stimulus is bounded by a 5 ms raised-cosine-envelope fade in and fade out as suggested by the International Telecommunication Union (2015).

Chapter 3

EXPERIMENTAL DESIGN

3.1 Italian phonology

In tonic (stressed) syllables, Standard Italian distinguishes seven vowel phonemes: the high front unrounded /i/, the close-mid and open-mid front vowels /e/ and /ɛ/, the low central /a/, the close-mid and open-mid back rounded vowels /o/ and /ɔ/, and the high back rounded /u/ (Bertinetto and Loporcaro, 2005). This inventory is organized along two principal articulatory–acoustic dimensions. The first dimension, tongue height, inversely corresponds to the first formant frequency F_1: /a/ occupies the highest F_1 position (maximally open jaw), whereas /i/ and /u/ occupy the lowest (maximally closed jaw). The second dimension, tongue advancement, corresponds to the second formant frequency F_2: the front vowels /i/, /e/, and /ɛ/ have high F_2 values (forward placement of the tongue), while the back vowels /u/, /o/, and /ɔ/ have correspondingly low F_2 values as a consequence of the rear placement of the tongue (Peterson and Barney, 1952). Together, F_1 and F_2 form a two-dimensional acoustic space that reliably separates the vowel categories of most languages and serves as the primary input to vowel identity perception (D. H. Klatt, 1980).

The most phonologically significant feature of the Italian system is the contrast between close-mid and open-mid realizations within each mid-vowel pair: /e/ (close-mid, lower F_1) versus /ɛ/ (open-mid, higher F_1) in the front series, and /o/ (close-mid, lower F_1) versus /ɔ/ (open-mid, higher F_1) in the back series. These contrasts are distinguished acoustically by a single scalar difference in F_1: in the measurements of Ferrero et al. (1978), /e/ and /ɛ/ differ by 140 Hz in F_1 (350 Hz versus 490 Hz), and /o/ and /ɔ/ by 160 Hz (390 Hz versus 550 Hz). The open/closed distinction therefore reduces to a one-dimensional acoustic contrast within each pair, making it particularly sensitive to variation in jaw aperture and a salient feature to study in the context of perceptual category boundaries between adjacent vowels.

However, the seven phoneme contrast is not maintained in all prosodic positions. In unstressed syllables, the inventory reduces to five phonemes: /ɛ/ raises towards /e/ and /ɔ/ raises towards /o/, which neutralizes the open/closed opposition entirely (Bertinetto and Loporcaro, 2005). This positional asymmetry has a direct practical consequence on experimental design, as stimuli intended to probe the /e/-/ɛ/ and /o/-/ɔ/ contrasts must be contained in a tonic syllable: only the stressed position guarantees the phonemic contrast is active. For this reason, all stimuli in the present experiment are produced in carrier words with a tonic target syllable, as described in Section 3.6.

The acoustic properties of the Standard Italian tonic vowels were documented by Ferrero et al. (1978) from recordings of Florentine university students pronouncing disyllabic /ˈCV.CV/ carrier words. These measurements are used as the Standard Italian end of the perceptual continuum constructed in Chapter 2. The full set of F_1 and F_2 values is given in Table 3.1. As described above, /a/ has the highest F_1 value (780 Hz), compared to /i/ and /u/ at with lowest (290 Hz and 320 Hz respectively). The clear separation between the close-mid and open-mid pairs in both the front and back series can also be noted.

Table 3.1: Means and standard deviations of F_1 and F_2 in Hertz for the seven tonic vowels of Standard Italian, from Ferrero et al. (1978). Speakers were ten Florentine university students.

	/i/	/e/	/ɛ/	/a/	/ɔ/	/o/	/u/
F_1	290 ± 35	350 ± 45	490 ± 50	780 ± 45	550 ± 40	390 ± 60	320 ± 40
F_2	2310 ± 140	2050 ± 110	1950 ± 100	1430 ± 80	970 ± 40	870 ± 80	800 ± 100

Regional variation in mid-vowel realization is, nevertheless, a pervasive feature of Italian phonology. The seven tonic vowel inventory is present across most of the peninsula, but the phonetic values of F_1 and F_2 for the mid vowels, as well as the lexical assignment of individual words to /e/ versus /ɛ/ and /o/ versus /ɔ/, vary substantially by region. For instance, Bertinetto and Loporcaro (2005) show that speakers from Florence, Milan, and Rome each show different acoustic and distributional patterns. The regional dimension of mid-vowel variability is the

phonological motivation for the present study. In particular, the experiment attempts to determine whether speakers of the variety of Italian spoken in Todi, whose tonic vowel realizations differ systematically from the Ferrero et al. (1978) Standard Italian references for both F_1 and F_2, as demonstrated in Section 3.2, show correspondingly different perceptual prototypes in the F_1-F_2 vowel space. To investigate this, the seven-vowel tonic system is considered as both a linguistic object under study, as well as the acoustic basis for the formant synthesis system described in Chapter 2.

3.2 The Todi variety of Italian

Umbria is a landlocked region in central Italy to the south of Tuscany and the north of Lazio. Situated in the Apennine range, its geographical position also reflects its cultural and linguistic status as neither Northern nor Southern, but rather appertaining to what comparative dialectology deems the Central Italian group. This area encompasses Tuscany, Umbria, Marche, and Latium, with Rome treated as a partially distinct case, and is distinguished from the Northern and Southern extremes by the fact that its speakers are, according to Canepari (2026), all native speakers of Italian: their regional speech derives from the continuous evolution of the local Latin-descended dialects, rather than from a sustained contact with other substrates. The dialects of western and southern Umbria belong to the *Italia mediana* (middle Italy) type, a compact zone delimited to its north and west by the Rome–Ancona Line, the bundle of isoglosses running between Rome and Ancona via the Tiber and Chiascio rivers, which is distinguished from both the Tuscan-based standard and the upper southern varieties by characteristic vowel features including metaphonic raising of stressed mid vowels (Vignuzzi, 1997). For this reason, Central Italian varieties have, according to Canepari (2026), "inherited the appropriate timbres" for the open/closed mid-vowel distinctions more faithfully than most other regional varieties. Nevertheless, there are deviations from this general rule: to the East and South of the Central dialect zone, including parts of Umbria, /ɛ/ and /ɔ/ may approach /e/ and /o/ in certain phonological environments. This is a source of variation that the present study addresses through the speaker selection criteria described in Section 3.3.

The Italian of the hill town Todi, located in the south-western corner of the province of Perugia along the Tiber river and bordering the neighboring province of Terni, is the object of this study. Vignuzzi (1997) identifies, on the basis of Moretti (1987), a transition zone from the western to the central Umbrian dialect type in the area running between Scheggia and Todi, which places the community at the southern boundary of this continuum and thus within the nucleus of the 'middle Italian' dialect zone. Its use as a sample of the broader Umbrian speech community is further supported by its geographical centrality within the region. Furthermore, the broader prestige history of Umbrian speech, as (Canepari, 2026) notes, can be remembered by the use of speakers predominately from Umbria and non-Roman Lazio as pronunciation models by Italian broadcasting in the 1950s and 1960s, before stricter Florentine-based norms became institutionally codified in subsequent decades. This historical standing lends further motivation to the investigation of the vowel systems of particular Umbrian varieties and how they differ from Standard Italian.

The acoustic measurements reported in Table 3.2 are derived from recordings of eight native Italian speakers from Todi (three male, five female) produced in the carrier-word procedure described in Section 3.3. Steady-state formant frequencies for F_1–F_5 were extracted at the midpoint of each carrier-word using Praat, and group means were computed. Per-speaker means were considered across speakers within each sex group, following the analysis methodology of Section 3.4. The resulting values are compared to the Standard Italian reference present in Ferrero et al. (1978).

Two systematic acoustic differences between Todi and Standard Italian emerge from the data. The first is a uniform elevation of F_1. Across all seven tonic vowels and both speaker groups, the Todi measurements exceed the Ferrero et al. (1978) reference in F_1 without exception. For the combined sample, F_1 offsets range from +54 Hz for /i/ to +132 Hz for /a/. The male group shows differences of +4 to +84 Hz and the female group +74 to +176 Hz. Because F_1 varies inversely with tongue height and jaw aperture, a uniformly elevated F_1 indicates that Todi speakers

Table 3.2: Mean F_1 and F_2 (Hz) for the seven tonic vowels of Todi speakers (males $n = 3$; females $n = 5$; combined $n = 8$.)

	F_1 (Hz)			F_2 (Hz)		
Vowel	Males	Females	Combined	Males	Females	Combined
/i/	310 ± 8	364 ± 33	344 ± 38	2444 ± 164	2694 ± 65	2601 ± 164
/e/	407 ± 19	469 ± 25	446 ± 38	2204 ± 16	2486 ± 63	2380 ± 154
/ɛ/	494 ± 2	581 ± 79	549 ± 75	2078 ± 44	2262 ± 131	2193 ± 140
/a/	841 ± 28	956 ± 61	912 ± 77	1422 ± 41	1495 ± 45	1467 ± 55
/ɔ/	634 ± 27	648 ± 47	643 ± 39	973 ± 46	1067 ± 30	1032 ± 59
/o/	474 ± 10	524 ± 33	505 ± 36	833 ± 44	928 ± 45	892 ± 64
/u/	349 ± 17	394 ± 23	378 ± 30	670 ± 34	744 ± 27	717 ± 47

produce the entire tonic vowel inventory with a more open jaw configuration than the Standard Italian reference. At these magnitudes, the displacement along the primary perceptual dimension of vowel height substantially exceeds Flanagan's (1957) just discriminable difference of approximately ±3% of the formant value.

The second pattern concerns F_2. In this case, the direction of the Todi-Ferrero difference depends on the front–back class of the vowel. For the three front vowels, F_2 is substantially higher in Todi than in Standard Italian: the combined offsets are +330 Hz for /e/, +291 Hz for /i/, and +243 Hz for /ɛ/. In contrast, the back vowels /a/, /ɔ/, and /o/ show negligible or modest F_2 differences (combined offsets of +37, +62, and +22 Hz respectively), and /u/ shows a reversal: its F_2 is lower in Todi than in Standard Italian (−83 Hz; male speakers: −130 Hz). Higher F_2 in the front series indicates a more advanced tongue-body position, consistent with a more peripheral or tense realization of the front vowels. The lower F_2 for /u/ conversely points to a more retracted high back vowel. Taken together, the F_2 pattern suggests that the Todi dialect expands the perceptual front–back contrast of the vowel space relative to Standard Italian, particularly at the high front and high back extremes.

The larger absolute offsets in the female group are expected and do not necessarily reflect a distinct dialectal pattern: women's shorter average vocal tract length produces higher absolute formant values throughout, and the scaling effect is proportionally

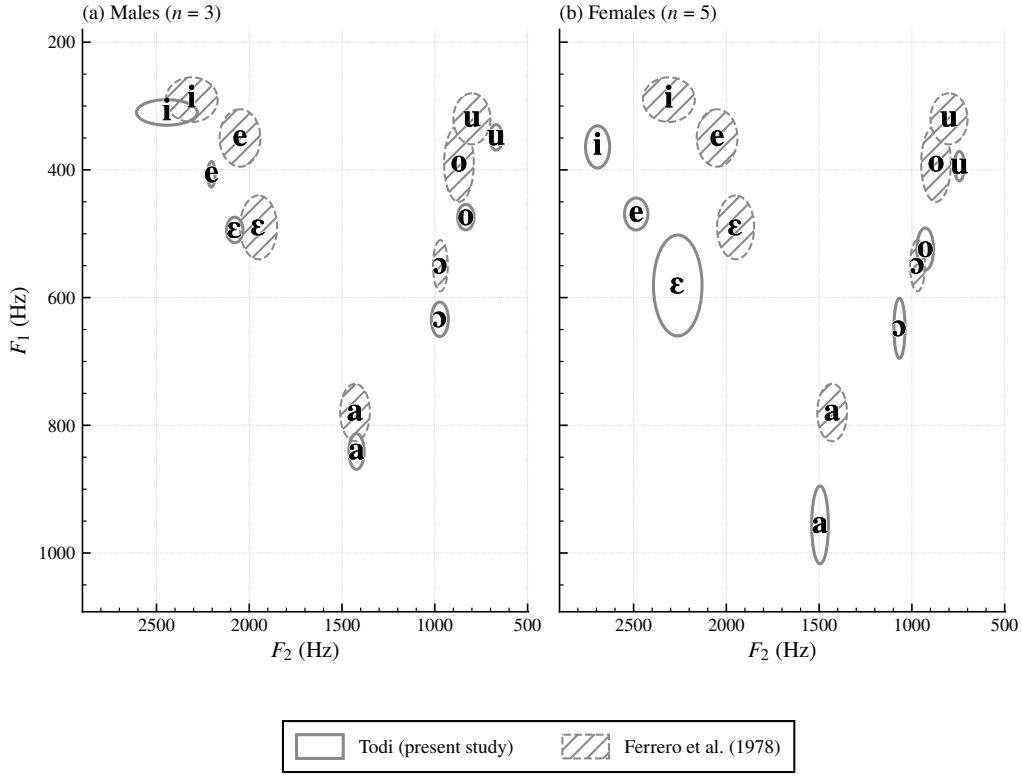

Figure 3.1: Average values in Hertz for vowels of male and female Todi speakers in the F_1-F_2 vowel space, with Ferrero et al.'s (1978) Standard Italian reference for comparison. Ellipses represent one standard-deviation from the mean.

greater on the F_2 axis (Rosner and Pickering, 1994). The directional peculiarities (elevated F_1, elevated F_2 for front vowels, depressed F_2 for /u/) are consistent across both sexes and are therefore interpreted as a genuine acoustic property of the Todi dialect rather than a sampling or experimental artifact. A limitation worth noting is that the Ferrero et al. (1978) Standard Italian reference does not report separate male and female values; the dialect–standard comparison is accordingly made against a mixed-sex reference. This constraint is acknowledged in Section 3.4 and revisited in Section 4.3.

3.3 Vowel data collection

Acoustic data for the Todi vowel space was collected using a stimulus word list modeled directly on that of Ferrero et al. (1978). The list consists of fourteen disyllabic Italian words, each containing one of the seven tonic vowels of Italian: /i/, /e/, /ɛ/, /a/, /ɔ/, /o/, /u/. The vowels were pronounced in a stressed open syllable preceded and followed by a single consonant (/ˈCVtV/ and /ˈCVdV/). Words are arranged as minimal pairs distinguished solely by the voicing of the intervocalic stop: the voiceless dental /t/ versus the voiced dental /d/ (Table 3.3). As in the original study, the initial consonants are labiodental or apico-dental in all tokens except /ˈputo/. In addition to the fourteen words, each speaker produced the seven vowels in isolation, which yielded twenty-one vowel productions per speaker in total.

Table 3.3: Stimulus word list from Ferrero et al. (1978).

Vowel	–t–	–d–
/i/	/ˈriti/	/ˈridi/
/e/	/ˈrete/	/ˈfede/
/ɛ/	/ˈlɛte/	/ˈsɛde/
/a/	/ˈrata/	/ˈrada/
/ɔ/	/ˈlɔto/	/ˈlɔdo/
/o/	/ˈvoto/	/ˈrodo/
/u/	/ˈputo/	/ˈsudo/

The use of an identical word list to Ferrero et al. (1978) serves two purposes. First, it provides a methodological anchor. The corpus consists of lexically meaningful Italian words, which reduces artifacts associated with the coarticulation of nonsense stimuli. The consonantal frame was selected by Ferrero et al. (1978) precisely because it does not produce systematic perturbations in mean vowel formant values. A statistical verification found no significant difference in F_1 or F_2 between the pre-/t/ and pre-/d/ contexts (Ferrero et al., 1978). Voicing does affect vowel duration: the vowel preceding /t/ is shorter than that preceding /t/ by 8–17%. Duration, nonetheless, is not the variable of interest here. Second, the identical stimulus design makes the Todi measurements directly comparable to the only published acoustic

description of the full Italian seven-vowel system that reports F_1–F_4 means, standard deviations, and coefficients of variation under controlled elicitation conditions. This comparability is exploited systematically in the present study.

These distinction between mid-vowels is effective only in stressed (tonic) syllables; in an unstressed position the system neutralizes to five vowels (Ferrero et al., 1978). Regarding the distributional properties of the mid vowels across Italian regional varieties, Canepari (2026) notes that the central-eastern strip of the peninsula, including Umbria, tends toward a more open realization of /e/ and /o/.

The assignment of each mid vowel word to the correct phoneme category was verified prior to data entry. The orthophonetic transcriptions provided by Ugoccioni and Rinaldi (2001). This step is necessary because Italian orthography does not mark the open/closed distinction; a word such as *rete* could in principle be pronounced with either /ɛ/ or /e/.

Speakers

Eight speakers from the Todi area of Umbria (Province of Perugia) were recorded in a single session in April 2026. The group comprised three males and five females, with ages ranging from 18 to 57 years (M = 29.4, SD = 14.9). Inclusion criteria required that each participant: (i) was born in or had spent their formative years in the Todi area; (ii) had Umbrian-born parents; (iii) reported no known speech or hearing disorder; and (iv) was free from illness on the day of the session. All speakers reported Italian as their first and daily language. Self-reported frequency of Umbrian dialect use varied from occasional to habitual across the group, as documented via a background questionnaire administered prior to recording (Table 3.4). One speaker (S8) was residing in Milan at the time of recording, with a total of ten years of residence outside of Umbria. Visual inspection of her individual vowel data revealed a distributional pattern that departed from the remaining seven speakers for /ɛ/. The formants measured for this vowel were much closer to the values reported in Ferrero et al. (1978). This departure is consistent extended residence in another linguistic environment. Clopper (2021) provides evidence that residential history is the primary determinant of dialect category representations, and that

Table 3.4: Biographic data of the eight speaker informants from Todi.

ID	Sex	Age	Birthplace	Residence	Umbrian parent(s)	Dialect (daily)
S1	F	23	Assisi (PG)	Todi (PG)	Both	3
S2	F	18	Città di Castello (PG)	Todi (PG	Both	2
S3	M	57	Foligno (PG)	Todi (PG)	Both (Todi)	3
S4	M	18	Marsciano (PG)	Todi (PG)	Both (Todi)	4
S5	F	28	Todi (PG)	Milan (MI)	Both	5
S6	M	18	Foligno (PG)	Todi (PG)	Both	4
S7	F	53	Todi (PG)	Todi (PG)	Both (Todi)	2
S8	F	20	Città di Castello (PG)	Todi (PG)	Both	2

mobile listeners' (those who have lived in multiple dialect regions) discrimination performance diverges from that of non-mobile native speakers. For this reason, S8's decade-long residence outside of Umbria could be considered extended cross-dialectal exposure. Nonetheless, if this were the case, a similar change among the other vowels would also be expected. However, this was not the case; the remaining 6 vowels showed means and standard deviations that matched those of the other 7 informants. For that reason, S8 was not treated as an outlier, and her measurements were included in the group statistics reported below.

Recording procedure

Recordings were made with a Tascam DR-40X portable digital recorder and an Audix HT5 slim-line headset condenser microphone powered via an Audix APS-911 battery-operated phantom power adapter. The headset configuration ensured a consistent microphone-to-mouth distance of approximately 2–3 cm across all speakers. This eliminated preventable variation in recording level due to microphone placement. Audio was captured at a sampling rate of 48 kHz with 16-bit precision. Each speaker read the stimulus words from printed instructions and was instructed to produce each item at a comfortable speaking pace. In the absence of an anechoic chamber, recordings were conducted in the best available acoustic environment, namely a room previously treated with sound-absorbing acoustic panels for sound-proofing. After completing the word list, each speaker produced the seven isolated vowels. No instructions about pitch or voice quality were given beyond the general

guidance to read naturally.

Formant extraction

Recordings were opened in Praat (Boersma and Weenink, 2024) and the steady-state portion of each target vowel was identified by visual inspection of the wide-band spectrogram. The steady state was defined as the central portion of the vowel interval, excluding consonant-to-vowel and vowel-to-consonant formant transitions, following the method described by Ferrero et al. (1978), who measured "the first four stationary formant frequencies ... at about the midpoint of each phone." Formant frequencies F_1–F_5 and corresponding bandwidths B_1–B_5 were read manually from Praat's formant display. The formants were calculated with Linear Predictive Coding (LPC) using the Burg algorithm and the standard settings for male and female speakers. Per-phoneme means were then computed by averaging across the set of steady-state measurements for each vowel

Manual extraction from the steady state was chosen over automated formant tracking for two reasons. First, it replicates the measurement procedure used by Ferrero et al. (1978), who also measured at stationary points on a spectrogram. This makes the comparison in Section 3.4 methodologically consistent. Second, the small corpus size (14 words plus 7 isolated vowels per speaker, across 8 speakers) made exhaustive manual inspection practical. The risk of undetected tracking errors that can affect automated extraction was thereby avoided.

Summary statistics

Formant and bandwidth values were aggregated across tokens and speakers to produce the grand mean and standard deviation per vowel category; all ten acoustic measures, F_1 through F_5 and B_1 through B_5, are reported in Table 3.5.

The standard deviations for F_1 and F_2 are broadly in line with those reported by Ferrero et al. (1978), which is interpreted as an adequate consistency for a sample of this size. The F_3 values range from 2594 Hz for /a/ to 3236 Hz for /i/. This is an expected result of the inverse relationship between F_3 and vocal-tract constriction location. Front vowels, produced with a palatal constriction, have higher F_3 than

Table 3.5: Per vowel mean and standard deviation for formant frequencies F_1–F_5 and bandwidths B_1–B_5.

	Vowel						
	/i/	/e/	/ɛ/	/a/	/ɔ/	/o/	/u/
Formant frequencies (Hz)							
F_1	344 ± 38	446 ± 38	549 ± 75	912 ± 77	643 ± 39	505 ± 36	378 ± 30
F_2	2601 ± 164	2380 ± 154	2193 ± 140	1467 ± 55	1032 ± 59	892 ± 64	717 ± 47
F_3	3236 ± 116	2912 ± 157	2887 ± 157	2594 ± 171	2918 ± 146	2794 ± 170	2825 ± 156
F_4	3979 ± 183	3984 ± 270	3952 ± 274	3778 ± 228	3655 ± 249	3634 ± 260	3681 ± 221
F_5	4767 ± 302	4617 ± 416	4632 ± 359	4672 ± 194	4498 ± 94	4431 ± 213	4419 ± 307
Bandwidths (Hz)							
B_1	108 ± 40	92 ± 44	86 ± 13	126 ± 29	188 ± 130	219 ± 138	320 ± 247
B_2	164 ± 111	177 ± 85	190 ± 88	157 ± 37	204 ± 141	243 ± 142	227 ± 100
B_3	489 ± 209	407 ± 137	395 ± 121	348 ± 91	322 ± 218	246 ± 65	453 ± 161
B_4	396 ± 202	493 ± 386	686 ± 628	622 ± 241	347 ± 71	302 ± 79	592 ± 350
B_5	403 ± 72	307 ± 109	396 ± 141	565 ± 72	171 ± 86	184 ± 200	523 ± 324

central or back vowels (Fant, 1960). The F_4 and F_5 values are comparatively stable across vowel categories, with ranges of 3634–3984 Hz and 4419–4767 Hz respectively. This is also expected for higher resonances, whose frequencies are less influenced by articulatory gestures.

Of the five bandwidth measures, B_1 and B_2 show the smallest standard deviations and therefore the most consistent measurements across speakers. B_1 for /u/ shows the highest group mean among the seven vowels (320 ± 247 Hz), but the standard deviation is nearly as large as the mean. This may indicate high between-speaker variability rather than a consistent acoustic property. Fant (1960) notes that lip rounding narrows formant bandwidths by reducing the radiation contribution. The elevated mean may therefore partly reflect measurement uncertainty at the low first-formant frequency of this vowel. B_3, B_4, and B_5 have substantially higher between-speaker variability, particularly B_4 for /ɛ/: 686 ± 628 Hz. Once again, in this case the standard deviation is comparable in magnitude to the mean. Nonetheless, all bandwidth values are used in the synthesis pipeline as filter parameters.

3.4 Vowel space analysis

Fig. 3.2 shows the seven Todi vowels in Bark F_1–F_2 space. The Bark scale approximates equal perceptual steps across the auditory frequency range and is used for inter-vowel distance comparisons; the psychoacoustic motivation for this choice is discussed in Section 3.6. Fig. 3.3 shows the female and male groups separately.

All three plots display the expected arrangement for the seven-vowel Italian system. Each phoneme occupies a distinct area in the F_1–F_2 Bark space. This effect is more pronounced for the Todi vowels than the Standard Italian vowels, which present some overlap between /o/ and /u/. The seven-way opposition further supports the conclusion that the Todi data was measured accurately. The male plot shows the same category structure at somewhat lower absolute Bark values. This is consistent with the longer vocal tracts and correspondingly lower resonant frequencies of male speakers (Fant, 1960).

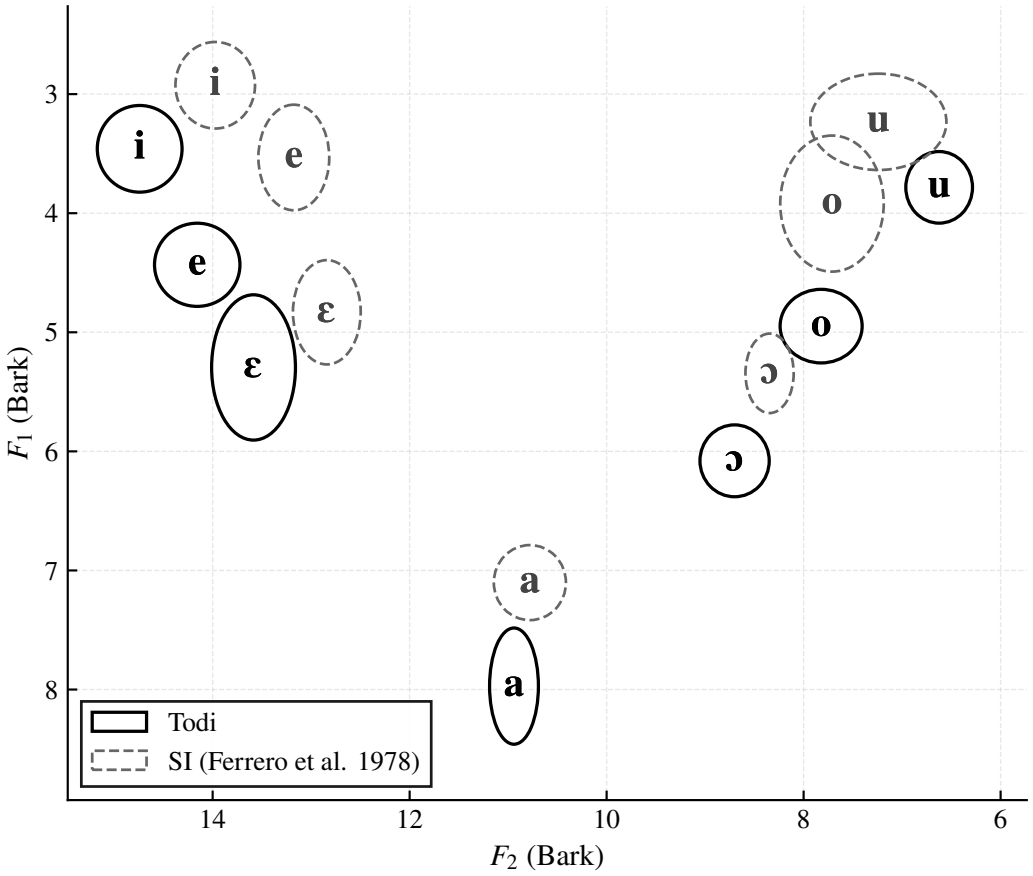

Figure 3.2: Vowel space in Bark for all speakers. Each ellipse spans ±1 standard deviation around the speaker-averaged mean.

Comparison

The Todi values are uniformly higher than the Ferrero et al. (1978) values across both F_1 and F_2 for all seven vowels. The primary explanation is the difference in gender composition between the two samples. Ferrero et al. (1978) report only that their informants were "10 university students, born and living in Florence with florentine parents," with no division by age or sex. The Todi sample contains five female and three male speakers. Because female speaker have anatomically shorter vocal tracts, their resonant frequencies are systematically higher than those of male speakers producing the same vowel quality (Johnson and Sjerps, 2021). A female-skewed

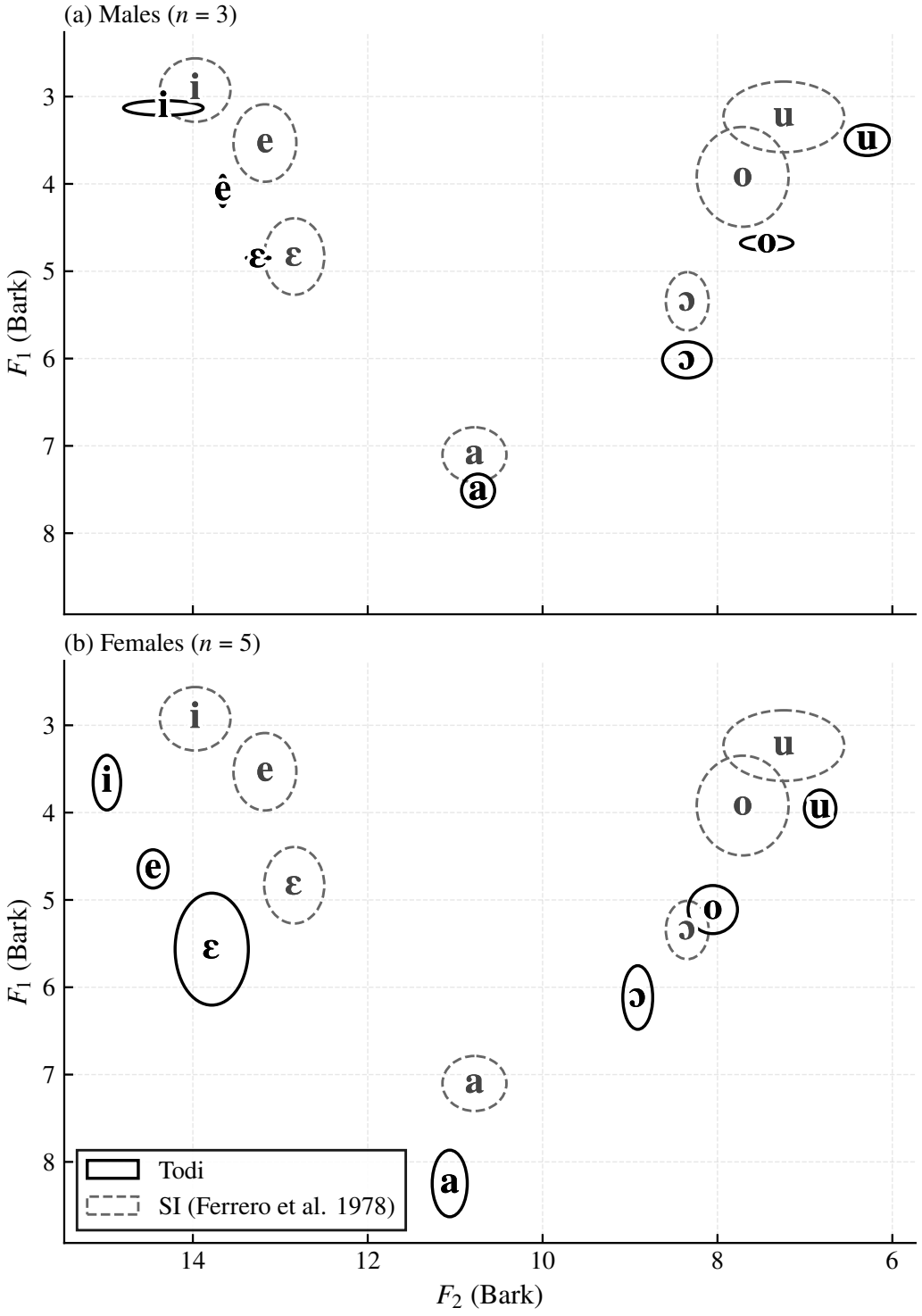

Figure 3.3: Vowel space in Bark for male and female speaker groups. Each ellipse spans ±1 standard deviation around the speaker-averaged mean.

sample will therefore produce higher group means. This may have contributed to the higher position of the Todi vowels with respect to the Standard Italian ones. It should be noted, however, that when the size of the groups were normalized (M = 3, F = 3), the relative positions of the means remained constant. The three retained female speakers were S1 (age 23), S2 (age 18), and S7 (age 53), selected to match the mean age of the male group as closely as possible (female mean: 31.3 years; male mean: 31.0 years); S5 and S8 were excluded. Therefore, it is assumed that the differences in the Todi vowels arise from regional pronunciation differences, rather than from sampling error. The modest reduction in the contrast between front vowels /e/ and /ɛ/ reflects the inclusion of S8.

Beyond gender and regional differences, the two datasets were collected nearly fifty years apart using different equipment and measurement methods. Ferrero et al. (1978) measured formants by hand from spectrograms produced by a Voice Identification Spectrograph 700. The present study uses the LPC algorithm to extract vowels from spectrograms in Praat. The methods are methodologically comparable, as both assume knowledge of the correct procedure to gather measurements from the speech spectrogram. The present study assumes that speaker habits have not departed radically since the measurements of Ferrero et al. (1978) were collected.

3.5 Theoretical framework

The foundation of the present methodology derives from the important work done by researchers in both signals engineering and its applications to psycho-acoustics over the course of the last three quarters of a century. As soon as synthetic vowels and consonants could be produced computationally with parametric models, thus allowing full control over the salient features being examined, the correlation between speaker production and listener identification could be tested and compared. As speech signals vary considerably from speaker to speaker even within the same language, any comparison of vowels must take into consideration the large range of variables (anatomy, vocal tract length, and so on) at play. Furthermore, the two facets of the cognitive processes behind speech, one motor and the other auditorily driven, are surprisingly disconnected: that is, what speakers think they are producing is not

necessarily perceived that way by listeners.

Continuous perception and category structure

Much research has gone into the categorization of speech sounds, in particular vowels and stop consonants. For occlusives (stop consonants), research has shown that listeners are able to perceive the difference between different phonetic categories, but are "deaf" to intra-category differences (Pisoni, 1973). This would appear to be a typical feature of most consonantal perception, however, vowels appear to function in a much different way. Steady state vowels, much like non-speech sounds, have been shown to be continuously perceived, meaning differences within and between categories are accessible for listener evaluation. The length of vowels, in particular in connected speech, can vary radically, from 250 ms in controlled laboratory settings, to less than 50 ms in spontaneous speech. In either case, the acoustic cues (formants) that characterize vowels are ever present, and research has shown that both vowels of short and long durations are perceived continuously. The categorization of vowels is therefore not absolute, and the discrimination of vowels within-category is considerably better than that of consonants: well above chance.

Many human perceptual systems that categorize various stimuli have been shown to have varying degrees of internal structure. When, within the same category, certain stimuli are preferred over others, category organization or hierarchy can be posited. The "goodness" or prototypicality of a stimulus is thus a privileged status as exemplars of a particular category. In human infants, phoneme categories have been suggested to be represented by prototypes at just 6 months of age (Goudbeek et al., 2005). Research shows that category membership is graded, and that within categories, the prototypes or best instances are particularly salient in perception (Kuhl, 1991). This effect has been called the perceptual magnet effect: that is, members surrounding a particular prototype are perceptually assimilated to it to a greater degree than the real psycho-physical distance. Prototypes thus anchor categories and provide a strong strengthening effect to the overall cohesiveness of the category itself. Interestingly, ratings for vowels would appear to be uniformly symmetric around a particular location, the category prototype. The consistency among

adults to subjectively evaluate vowel sounds using an internal mental prototype as a reference is compelling evidence supporting studies in vowels across dialects. In summary, due to the perceptual magnet effect, vowel prototypes attract neighboring stimuli. Goodness ratings peak at the prototype center and decline symmetrically the farther stimuli move from its "hot spot."

Phonetic targets and the hyperspace effect

As mentioned, speech can range from a variety of realizations, from the most precisely crafted laboratory sentences (hyperarticulated speech) to spontaneous, casual utterances (hypoarticulated speech). While theoretically pleasing, these two extremes are, in practice, experimentally difficult to elicit (Johnson, Flemming, and Wright, 1993). Nonetheless, the importance of hyperarticulated speech to phonetics research cannot be overstated. Clear, highly articulated speech carries the most amount of information possible: reduction processes, while reducing effort, in terms of articulatory gesture, also reduce the information content of the speech. Casual speech can thus be seen as temporal overlap between articulatory gestures: the gestures themselves, the targets, are not deleted, but rather are rather prevented from making an appearance acoustically. Thus, while gestures are undershot, covered, or blended together, research shows that they still possess precise mental targets. In order to explore listeners' phonetic targets, Johnson, Flemming, and Wright (1993) developed the Method of Adjustment (MOA), which consists in using a computer to manipulate one or more parameters of a speech synthesizer until it correctly pronounces a speech sound. By using the MOA, that is, a single synthetic voice, the problem of personal speaker variation is avoided and a listener's perceptual expectations for vowel sounds in a particular language/dialect can be assessed, thus allowing cross-linguistic phonetic comparisons. This methodology showed great consistency across trials, although it was shown that the vowel space of the ideal setting for the speech synthesizer was expanded relative to measured (produced) formant values of vowels. This expansion, the so-called "hyperspace" effect, beyond showing the disparity between discrete mental categories and production, also highlights that phonetic targets are hyperarticulated. Furthermore, it has been shown that speak-

ers of the same dialect answer in the same way regardless of previous experience, training, or instruction in phonetics, in particular when they are asked to evaluate a sound as they would say it.

By placing sounds along a continuum of values, it is thus possible to allow speakers to adjust, or select, the sound on the continuum that most accurately represents how they say the sound, and thus arriving at their mental prototypes. Thus for two dialects that are assumed to be different in their realizations of the same, or a subset of, vowels, one would expect to find the speakers of two different dialects to have unique sets of locations of mental prototypes for the phonemes that are divergent.

Dialectical perception

Dialect classification relies on two core abilities: recognizing the sounds of speech and understanding their social meanings. Research shows that people with limited hearing or difficulties in social processing do less well at these tasks than adults with typical abilities (Clopper, 2021). This happens because listeners must connect specific sound patterns they hear to broader social groups, such as the region a speaker comes from. These connections form the basis for deciding where someone is from. People learn the connections through their own lifetime of language exposure and through shared cultural ideas about how people from different places speak. As a result, how accurately someone can classify dialects varies widely from person to person.

Adults generally perform these tasks with greater accuracy than children, especially when they hear their native dialect or relatively standard forms of a language. Both phonetic processing and dialect identification improve with greater familiarity, are shaped by what the listener expects, and are influenced by social stereotypes. This is why listeners can often tell that an unfamiliar way of speaking comes from outside their own area, yet they may struggle to name the exact place it comes from. They may simply register the speech as “different from mine, so not from here.” In short, a person’s linguistic history and social experiences play a central role in how well they identify dialects. Listeners therefore hold the clearest mental pictures of their own native variety and classify speakers from their home region more accurately

than others. Studies also find that performance tends to be stronger for varieties spoken over smaller geographic areas than for those spread across larger ones.

The advantage of knowing one's native variety extends to people who have moved around. Listeners who have lived in several different dialect regions show improved classification not only for their original home variety but also for the varieties of every place they have resided. This accumulated experience gives mobile listeners an overall edge compared with people who have stayed in one place.

Several challenges arise when researchers try to measure how dialects vary and how people perceive them. First, if a listener does not know or cannot guess where a new speaker is from, the listener cannot reliably link the speaker's distinctive speech features to a specific region, even when those features clearly signal regional background. Second, ordinary people without linguistic training do not notice every detail of regional variation that trained sociolinguists document. Although non-experts readily hear that dialects differ, the mental categories they form may not match the categories linguists use. Third, the ability to connect speech sounds with social and geographic information develops gradually. Young children already show adult-like skill at telling sounds apart, but they lack enough knowledge about places and social groups to classify dialects like adults do. Fourth, each listener's mental map of dialect differences is shaped by the actual range of variation they have encountered; an objective count of how different two dialects are does not always predict how easily people can tell them apart. Finally, and most important for the present work, the amount of time a person has spent living outside their original dialect region strongly affects how they perceive and process speech.

Several of the problems outlined are not relevant to the current study because only adults without social or auditory impairments participated. Adult listeners possess fully developed social and geographic knowledge, eliminating the developmental limitations seen in children. In addition, the study focuses on perceptual classification of vowel features on a continuum rather than open-ended geographic labeling of unknown speakers. Participants' own mobility history (years lived outside Umbria) is explicitly measured and treated as a key variable, directly addressing the influence

of lifetime geographic experience on dialect perception. Because virtually all adults in contemporary Italy have had extensive exposure to standard Italian from an early age, this factor does not vary meaningfully across participants and is therefore not included as a variable in the analysis.

Primary vowel dimensions

The importance of a speaker's dialectal background in shaping vowel pronunciation became clear soon after the sound spectrograph appeared, nearly seventy-five years ago. The most influential study from that period remains the work of Peterson and Barney (1952) at Bell Telephone Laboratories. They recorded ten American English vowels produced in a /hVd/ frame by men, women, and children, then measured the frequencies and amplitudes of the first three formants along with the fundamental frequency. Their data revealed that steady-state formant patterns serve as the main acoustic information listeners rely on to recognize vowels. At the same time, the study documented wide differences across speakers, even within the same gender or age group. These differences arise from variations in human anatomy, such as vocal-tract length and shape. Formant values also shift as a result of different speaking rates and phonetic contexts, which further complicates an idea model that directly maps acoustic features to perceived vowel identity.

When listeners in Peterson and Barney's (1952) experiments tried to identify vowels, overall accuracy proved lower than expected, with most errors occurring between neighboring categories on the vowel chart. In other studies, it has been shown that untrained participants often struggle not because they fail to hear quality differences, but rather because they have difficulties in translating those qualities into a discrete set of orthographic labels.

In summary, individual speaker differences obscure the ideal mental targets listeners actually use. For this reason, the MOA developed by Johnson, Flemming, and Wright (1993) is particularly suited to the purposes of the present work.

3.6 Continuum design

Research from the early 1950s examined the question of which spectral properties of vowels carry the most weight for listener identification. Delattre et al. (1952) produced synthetic steady-state vowels and systematically manipulated their formant frequencies, establishing which spectral configurations produced the highest identification rates for each vowel category (Raphael, Borden, and Harris, 2011).

The concept of the *formant* is central to research in this domain. A formant is the acoustic manifestation of a resonance peak in the vocal-tract transfer function (Fant, 1960). For the acoustic characterization of vowel sounds, the relevant parameters are the center frequencies of the three or four lowest resonances, referred to collectively as the *formant pattern* (Raphael, Borden, and Harris, 2011). Perceived vowel quality is governed principally by the formant pattern; the relative amplitudes and bandwidths of the individual formants contribute only modestly to vowel identity.

The principal finding of Delattre et al.'s (1952) experiments was that listeners could, in the great majority of cases, correctly identify a vowel on the basis of the first two formants alone. A further asymmetry exists between front and back vowel classes: whereas front vowels required both F_1 and F_2 for reliable identification, back vowels could be identified from a single formant whose frequency fell between the natural values of F_1 and F_2 (Raphael, Borden, and Harris, 2011). This asymmetry points to a differential perceptual role for F_3: its contribution to perceived vowel quality is appreciably greater for front vowels than for back vowels. The primacy of F_1 and F_2 in vowel identification is thus a foundational result of acoustic phonetics and is corroborated by subsequent large-scale listener studies (Hillenbrand et al., 1995).

Higher formants

As mentioned briefly in Section 2.4, for the synthetic vowels used in the present study F_3 is held constant at the Todi speaker value throughout; F_4 is fixed at the mean of the Todi and Ferrero et al. (1978) values; and F_5 is fixed at 4500 Hz. Bandwidth values are taken from the Todi speaker measurements. The latter two decisions are motivated by the absence of bandwidth and F_5 data in Ferrero et al.'s (1978) measurements; F_5 is therefore set to the theoretically expected value for an

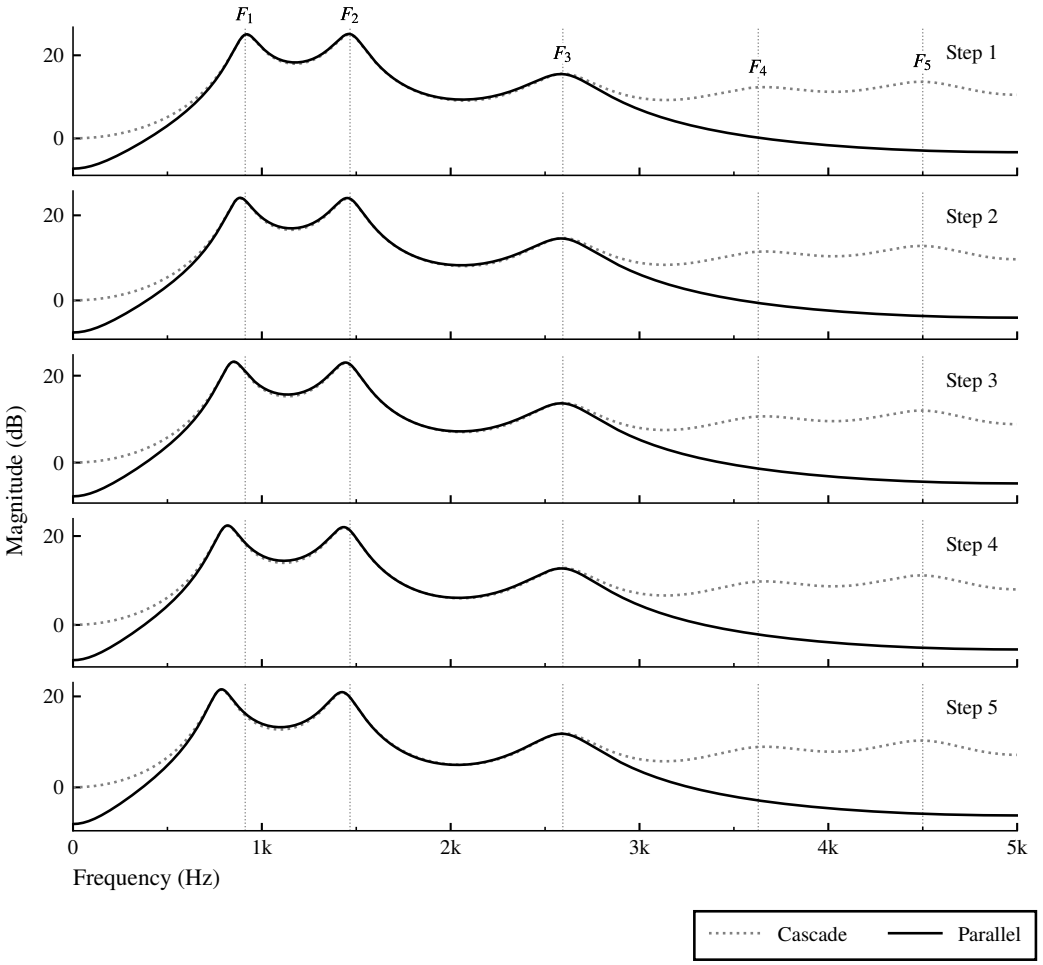

Figure 3.4: Variation of F_1 and F_2 on the /a/ continuum. The formants vary from the Todi target (Step 1) to the Standard Italian target (Step 5).

adult male with an average vocal-tract length (Fant, 1960). On the basis of the acoustic evidence reviewed above, the effect of holding bandwidths and F_5 constant is expected to be perceptually negligible.

Accordingly, vowel perception experiments typically allow F_1 and F_2 to vary while fixing the higher formants (F_3–F_5) at reference values. This reduces the experimental variables to the dimensions under investigation.

Psychoacoustic scaling

The human auditory system does not encode frequency linearly. As discussed in Section 2.4, the cochlea operates as a bank of bandpass filters arranged tonotopically, with frequency resolution that is finer at low frequencies and progressively coarser at higher ones. Distances in the F_1–F_2 plane are therefore more meaningfully expressed in units that reflect the critical-band structure of auditory perception than in absolute Hz values. The *Bark scale*, a psychoacoustic frequency scale, approximates the spacing of the 24 critical bands of the human auditory system and has been widely employed in psychoacoustics and speech perception research (Raphael, Borden, and Harris, 2011). Evidence from speech-sound studies indicates, moreover, that tonotopic distances between spectral peaks are fundamental to the perception of phonetic quality (Traunmüller, 1990). Two practical considerations motivate the use of Bark here. First, it enables the definition of perceptually uniform step sizes across the F_1–F_2 plane. Second, the conversion formula of Traunmüller (1990) is computationally straightforward, with a stated accuracy of ±0.05 Bark across the range of vowel formant frequencies characteristic of adult men, women, and children.

Formant values in Bark thus permit linear measurement of the perceptual distance between each vowel pair. For each continuum, listeners are asked to select the step that best matches their ideal mental pronunciation of each vowel. A continuum is therefore constructed for each of the seven vowel pairs. Its design follows two criteria: step sizes must respect the just discriminable difference for formant frequency, and the total number of steps must remain small enough to limit listener fatigue. According to Flanagan (1957), the difference limen for formant frequency

is approximately ±3% of the formant center frequency; at 400 Hz this corresponds to approximately 0.22 Bark, and at 2600 Hz to approximately 0.39 Bark. On the basis of the Bark distances reported in Table 2.2, a five-step continuum between the dialect and Standard Italian endpoints provides adequate perceptual resolution while keeping listener fatigue manageable; the resulting Bark step sizes for each vowel pair are given in Table 2.2.

To prevent participants from using stimulus position as an implicit response cue, the present study adopts the procedure of Johnson, Flemming, and Wright (1993): the orientation of each continuum is randomized from trial to trial. On 50% of trials the dialect endpoint appears on the left and the Standard Italian endpoint on the right; the remaining 50% present the reverse arrangement.

To assist participants in linking the IPA symbol displayed for each vowel to its phonetic referent, an Italian carrier word containing the target vowel in the stressed syllable is shown alongside each response button. These words are drawn from the exemplar set used for phonetic illustration in the *DÍPiN* (Canepari, 2026), and are considered sufficiently unambiguous for this purpose, although the method is not entirely free from complications arising from regional differences in lexical phonology and is a limitation to this methodology.

3.7 Task 1: Synthesis quality evaluation

Evaluating synthetic speech is a difficult task. There is no single answer to the question of how or what to evaluate. Much synthesis research has proceeded without any formal evaluation at all, or an ad-hoc evaluation was used (Taylor, 2009). The result is that testing in synthetic speech systems is not a simple or widely agreed upon area. It therefore follows that when interested in testing these systems for practical purposes, we can follow the testing and evaluation procedures used in mainstream engineering, as this area is particularly rich in the area of testing real systems. To that end, the user requirements are the first thing to consider. Obviously, for our purposes, and generally, the two main goals of synthetic speech systems are to be intelligible and natural.

Naturalness can be tested by playing some speech and asking listeners to simply rate what they hear. Often this is accomplished with a simple Likert scale from 1 to 5, where 1 is bad and 5 is excellent, known as mean opinion score (MOS) testing. One key problem of this method is that listeners, rather than measuring naturalness, tend to report how much they like the system (Taylor, 2009). This can be seen especially in tests where real speech is included along with synthetic speech: if naturalness were being assessed all real speech would achieve a perfect 5 rating, but this is not the case. "Pleasant" sounding voices often outscore voices with less desirable or aesthetically pleasing features. Furthermore, comparison between tests in different times and locations is not an absolute score of the systems performance and are thus not comparable.

A more robust system for measuring speech quality is given by a method called MUSHRA, or "MUlti Stimulus test with Hidden Reference and Anchor," released by International Telecommunication Union Radiocommunication Sector (ITU-R). The MUSHRA method has been successfully tested and is suitable for evaluation of intermediate audio quality and gives accurate and reliable results (International Telecommunication Union, 2015). MUSHRA was designed to assess audio systems that introduce noticeable impairments, such as those found in internet streaming or digital broadcasting, where bandwidth limits often force compromises in quality. The method follows a double-blind, multi-stimulus format. In each trial, listeners hear short audio excerpts and can switch between the versions of the same material: the open reference (unprocessed original), hidden reference, and processed signals under test. Listeners rate each signal on a continuous quality scale from 0 to 100, divided into five labeled intervals: excellent (80–100), good (60–80), fair (40–60), poor (20–40), and bad (0–20). Because they compare all signals directly, they naturally sort and rank them before assigning scores. Furthermore, the MUSHRA guidelines describe clear statistical procedures for analyzing the data.

For the present experiment, the effect of jitter, shimmer, and tremor will be considered in the context of the synthetic vowels and a MUSHRA-inspired task will be implemented that retains the continuous quality scale and the five labeled intervals,

but departs from the standard in several deliberate respects. The standard anchors (low-pass filtered versions of the stimulus at 3.5 kHz and 7 kHz) are omitted, as those anchors are entirely unrelated to the effects under study. This change was selected in order to minimize participant fatigue from an abundance of listening conditions. With seven vowels across three different conditions (natural, robotic, rough), there are a total of 21 conditions to evaluate, which should require a few minutes to complete.

Figure 3.5: The MUSHRA interface designed for the experiment. The layout matches the example specified in ITU-R BS.1534-3.

While the full ITU-R protocol specifies mandatory anchors and a hidden reference, several well-established variants of the MUSHRA methodology omit one or both of these elements, and their use is increasingly documented in the speech synthesis literature. Varadhan et al. (2025) provide a systematic critique of standard MUSHRA, identifying two primary shortcomings that are directly relevant to the present experiment: reference-matching bias, whereby listeners unconsciously align their ratings with the reference rather than assessing each stimulus independently, and judgment ambiguity, which arises from the absence of fine-grained rating criteria. They demonstrate that the ITU-R anchor introduces its own distortion artifacts

and can unfairly suppress scores for well-synthesized stimuli that bear no resemblance to codec-type degradation. They conclude that "there is merit in conducting MUSHRA evaluations without anchors". Showing the reference openly, rather than embedding it as a hidden item, further addresses reference-matching bias by giving listeners an unambiguous quality target from the outset, a design adopted in recent TTS evaluations (Perrotin et al., 2023). The present experiment therefore follows a MUSHRA-inspired naturalness rating task that retains the 0–100 continuous quality scale and five labeled intervals of ITU-R BS.1534-3, while omitting the standard anchors and presenting the reference explicitly. This is consistent with current practice in the speech synthesis evaluation literature.

To summarize, the experiment uses three synthesis conditions, all produced by the same parallel formant synthesizer with an LF glottal source, differing only in the degree of perturbation applied to the source signal. These effects are as follows: **Robotic**: no jitter, shimmer, or tremor. This is a perfectly periodic source, producing the characteristic flat, mechanical quality of unperturbed synthesis. **Natural**: moderate jitter (0.5%), shimmer (1%), and 4 Hz sinusoidal tremor (2%). These values were chosen to emulate a human sounding voice. **Rough**: high jitter (2%) and shimmer (5%). These values produce a rougher vocal quality. Listeners rate each condition on the quality/naturalness scale against the natural speech reference. The three conditions thus span a perceptual range from over-regular to over-perturbed, with the natural condition hypothesized to score the highest.

3.8 Task 2: Forced-choice identification

The identification task addresses whether the synthesized Todi dialect vowels are reliably categorized by listeners into the expected phonemic categories of Italian. Vowel perception has long been analyzed via confusion matrix, in which each stimulus is associated with a probability distribution over possible responses (Hillenbrand et al., 1995). This approach quantifies both accuracy and systematic misidentification. Peterson and Barney's (1952) study demonstrated that English vowels occupy overlapping regions of the F_1–F_2 plane, making some degree of cross-category confusion inevitable. Hillenbrand et al. (1995) replicated and extended these findings

with a larger and more demographically varied speaker sample. This study confirmed that overlap is a stable property of the vowel space rather than a measurement artifact.

The present task adopts the same procedure. The identification stimuli are the seven Todi endpoint synthesized vowels: /i/, /e/, /ɛ/, /a/, /ɔ/, /o/, /u/. The synthesized vowels use the Todi targets, that is, step 1 of each continuum (the formant values taken directly from the Todi speaker averages). The task measures identification of the dialect phonetic targets in isolation; Ferrero et al.'s (1978) Standard Italian measurements are not compared.

Figure 3.6: The forced-choice identification task interface, showing full list of example words. Participants were allowed to play the sound twice.

Each vowel is presented three times, yielding 21 trials in total. The use of multiple repetitions per item follows standard practice in identification experiments and provides a more stable estimate of each listener's response probabilities. This is necessary for constructing a well-populated confusion matrix with a small number of stimuli. Trials are presented in a randomized order, unique to each participant, to prevent order effects and sequential response biases. On each trial, the listener clicks a button to hear a single vowel and is then presented with a response panel

displaying the seven vowel options as clickable buttons. Each button is labeled with the IPA symbol.

A collapsible help box with the full list of Italian carrier words is available throughout the task. The carrier words serve as a lexical anchor for the phoneme category and are drawn from the same set used in the prototype selection task. This ensures consistency across the experiment. The listener may play the vowel sound only twice, but must play it at least once before responding. They cannot proceed until a response is selected. Reaction time from stimulus offset to response is logged alongside the response itself.

The experimental output is a 7×7 confusion matrix. Rows index the stimuli and columns index listener responses, aggregated across repetitions and participants. Rows and columns are normalized to conditional response probabilities to produce a row-stochastic matrix. Overall percent correct is computed as the mean of the diagonal. Off-diagonal patterns are interpreted in terms of the acoustic proximity of vowel pairs in the F_1–F_2 plane. Confusions between vowels that are close in Bark space are expected, whereas unexpected confusions may indicate failures of the synthesis to render the relevant acoustic cues adequately. The identification data are also examined in relation to listener group (Todi vs. Standard Italian).

Following the reasoning of Johnson, Flemming, and Wright (1993), it is predicted that listeners' selections will approach the hyperarticulated targets of their own dialect, such that a vowel that is peripheral in Standard Italian may be perceived as prototypical by a Todi listener and vice versa. If this is the case, Umbrian listeners would show higher identification accuracy for the Todi stimuli than Standard Italian listeners, whose category prototypes would be assimilated by the nearest Todi targets, reducing accuracy. Observed differences in the diagonal of the confusion matrix between the two groups are therefore interpreted as evidence for dialect-specific vowel targets rather than a failure of the synthesis.

3.9 Task 3: Prototype selection

The prototype selection task investigates where, along the perceptual continuum from the Todi to the Standard Italian vowels each listener's phonemic prototype is located. Kuhl's (1991) perceptual magnet effect proposes that vowel categories are not uniformly structured but are organized around prototypical exemplars that function as attractors. Tokens near the prototype are perceptually assimilated toward it, while tokens near the category boundary are more discriminable. Since speakers from Todi and Standard Italian speakers have been exposed to different vowel distributions, the perceptual magnet for any given phoneme is predicted to be located at different points along a dialect-to-standard continuum for the two groups. Assuming listeners' phonemic targets are hyperarticulated relative to the acoustic mean of their input, the perceptual prototype should lie at or beyond the measured values.

Figure 3.7: The prototype selection task interface, showing the reference word and continuum from 1 to 5.

It is therefore predicted that listeners from Todi will select continuum steps closer to the Todi variety target, while Standard Italian listeners will select steps closer to the measurements collected in (Ferrero et al., 1978). The magnitude of the group difference is predicted to vary across vowels according to how acoustically distinct the two endpoints are.

The continuum as a method for probing perceptual category structure has a well-established precedent in the categorical perception literature (Pisoni, 1973). The fact that listeners discriminate between-category pairs more accurately than within-

category pairs implies the existence of a boundary whose location differs predictably across listener groups. The present task does not measure discrimination, but instead asks listeners to directly identify the step that best represents their own pronunciation, as implemented by Johnson, Flemming, and Wright (1993).

The stimuli for each vowel are the five steps of the Bark continua described in Sections 2.4 and 3.6. Step 1 corresponds to the Todi target, and step 5 corresponds to the Standard Italian target. F_1 and F_2 are interpolated linearly in Bark space at equal intervals using Traunmüller's (1990) formulae, ensuring that equal physical steps along the continuum correspond to approximately equal perceptual distances. The resulting Bark distances between targets and their step sizes are reported in Table 2.2.

The task is presented once per vowel for a total of seven trials in total. In each trial, all five continuum steps simultaneously presented to the participant as labeled playback buttons from 1 to 5 arranged horizontally on the screen. Listeners are required to play every step before a response becomes available: the selection buttons are disabled until all five steps have been heard at least once. The prototype selection judgment is only meaningful if the listener has sampled the full range of variation on the continuum, therefore listeners may replay any step as many times as they wish after the initial mandatory listening pass. As in Johnson, Flemming, and Wright (1993), the direction of the continua are randomized in each trial so that 50% of the time the continuum, from left to right, begins at the Todi target and ends at the Standard Italian target. The other 50% of the time the inverse is true. This ensures that participants listen carefully and are not biased by the direction of the continua.

Once all steps have been heard, a button appears beneath each step. The listener selects the step that most closely matches their own pronunciation of the target vowel, then confirms their choice. As in the forced-choice identification task, a help panel displaying the IPA symbol and carrier words for the target vowel is available throughout the trial and can be toggled on or off. Trial order is randomized across participants. The task is administered after the identification task so that participants have already demonstrated familiarity with the seven vowel categories before being

asked to locate their prototype on a continuum.

The output for each trial is an integer from 1 to 5 that corresponds to the step. A full log of playback events and response time is also recorded.

3.10 Participants

Forty-eight participants began the experiment. Two were excluded due to no data being recorded, and eight were subsequently excluded because their total session duration fell below seven minutes, the minimum duration considered sufficient for genuine task completion, yielding a final sample of 38. Participants were assigned to one of two groups according to the region in which they spent the majority of their childhood: the Todi group (n = 16), and the non-Umbrian group (n = 22), comprising speakers from other Italian regions. To ensure geographic specificity within the Todi group, recruitment for that cohort was conducted through a dedicated distribution link circulated among personal and social networks to persons residing in the *comune* of Todi. The non-Umbrian group was recruited through a separate general link open to adult native speakers of Italian from any region. Participation was voluntary and unpaid in both cases.

Inclusion required that participants be native speakers of Italian with no self-reported hearing impairment, complete the experiment on a desktop or laptop computer, and use headphones or earphones as specified in the instructions. Mobile devices were blocked from participating.

Prior to the experiment tasks, all participants completed a short background questionnaire comprising six items: age; sex; childhood region, recorded as a categorical selection from a list of Italian regions; years of residence outside Umbria, a numeric entry treated as a measure of dialect contact continuity for the Todi group (see Section 3.5); highest level of education; and self-reported dialect usage frequency, rated on a five-point Likert scale from 1 (never) to 5 (always).

The final sample comprised 20 women and 18 men (Todi: 12 F, 4 M; non-Umbrian: 8 F, 14 M). Age ranged from 19 to 75 years (M = 36.8, SD = 14.4); the Todi group (M = 33.8, SD = 16.6, range 19–75) showed a somewhat wider age spread

than the non-Umbrian group (M = 38.6, SD = 12.8, range 23–67), consistent with the community-based recruitment approach used for that cohort. The non-Umbrian group included speakers from 12 Italian regions, with Tuscany (n = 6), Piedmont (n = 4), and Campania (n = 4) most represented. Descriptive statistics for the full sample, broken down by group, are given in Table 3.6.

Table 3.6: Participant demographics by group. Years outside Umbria is reported for the Todi group only, as it is not applicable to non-Umbrian participants.

Characteristic	Todi (n = 16)	Non-Umbrian (n = 22)	Total (N = 38)
Age: M (SD)	33.8 (16.6)	38.6 (12.8)	36.8 (14.4)
Age: range	19–75	23–67	19–75
Sex: female	12	8	20
Sex: male	4	14	18
Education: *diploma*	5	10	15
laurea triennale	3	9	12
laurea magistrale	9	8	17
dottorato	0	1	1
Dialect frequency: M (SD)	1.82 (0.98)	0.25 (0.63)	0.84 (1.09)
Years outside Umbria: M (SD)	3.9 (5.1)	—	—
Years outside Umbria: range	0–17	—	—

3.11 Ethics and distribution

Participation was entirely voluntary and unpaid. Before the experiment began, participants were presented with an information screen describing the purpose of the study, the nature of the tasks, and the expected duration (15–20 minutes). Informed consent was obtained digitally via a mandatory acknowledgment step; the experiment could not proceed without it. No personally identifying information was collected: participants were assigned anonymous numeric identifiers, and all stored data are keyed to these identifiers only. The demographic questionnaire collected age, sex, region of origin, and language background, which are do not constitute sensitive personal data. Data are stored encrypted on a password-protected server and will not be shared beyond the research team. The study was conducted in accordance with

general principles of research ethics applicable to perceptual studies as suggested by the Acoustical Society of America (2018).

Participants accessed the experiment via a standard web browser on any computer with audio output. On the opening screen they were told to use headphones or earphones, to set a comfortable listening volume, and to complete the session in one sitting in a quiet environment. The session then proceeded to the three tasks. In Task 1 (naturalness rating), participants evaluated seven vowels in three synthesis conditions each. In Task 2 (forced-choice identification), they heard one vowel token at a time and selected the closest matching vowel from seven buttons. In Task 3 (prototype selection), they listened to the five steps of a synthesis continuum and identified the step that best matched their own habitual pronunciation of each vowel.

The ordering of tasks was fixed for all participants: MUSHRA first, identification second, and prototype selection last. Within each task, vowel order was randomized per participant. Prior to the three tasks, participants completed the background questionnaire. Audio stimuli were pre-loaded before the task began to avoid playback delays during the trials.

The experiment was built using jsPsych, an open-source JavaScript framework for browser-based behavioral experiments (Leeuw, Gilbert, and Luchterhandt, 2023). It was served through JATOS, an open-source experiment management platform that handles participant routing, result collection, and data export (Lange, Kühn, and Filevich, 2015). The JATOS instance was deployed on a cloud server and accessed via a dedicated domain. Custom jsPsych plugins were written for the MUSHRA rating interface and the prototype selection task, as the standard jsPsych plugin library does not include these trial types. Trial-level response data (ratings, response times, play logs, and selected steps) were serialized as JSON and stored server-side after each completed task.

Chapter 4

RESULTS AND DISCUSSION

4.1 Task 1: MUSHRA

Task 1 employed a MUSHRA (Multi-Stimulus test with Hidden Reference and Anchor; International Telecommunication Union (2015)) test to assess the perceived naturalness of three synthesized stimuli conditions. All three were produced from the same LF glottal-source model (Fant, Liljencrants, and Lin, 1985) and static Todi vowel targets, differing only in source perturbation.

All seven tonic vowels were tested. Listeners rated each set of simultaneous conditions on the ITU-R Continuous Quality Scale (0–100), labeled *Pessimo* ("Bad", 0–20), *Scarso* ("Poor", 20–40), *Discreto* ("Fair", 40–60), *Buono* ("Good", 60–80), and *Eccellente* ("Excellent", 80–100). Following a standard MUSHRA non-engagement criterion (> 86% of responses at the default slider position), one participant was excluded, leaving a final MUSHRA sample of $N = 37$ (Todi: 16, Toscana: 4, Other: 17).

Results

Collapsed across vowels and listener groups, mean ratings fell entirely within the *Poor* category (20–40) for all three conditions (Table 4.1). An ordering of natural > robotic > rough can be observed, with means of 33.6 ($SD = 27.6$), 29.5 ($SD = 28.9$), and 23.8 ($SD = 22.6$), respectively. Medians were markedly lower. This indicates strong positive skew consistent with the high inter-listener variance documented in MUSHRA evaluations more broadly (Varadhan et al., 2025).

A Friedman test on per-participant condition means (collapsed over vowels) confirmed a significant effect of condition: $\chi^2(2) = 18.41$, $p < .001$. Post-hoc Wilcoxon signed-rank tests with Bonferroni correction ($\alpha = .017$ for three pairwise comparisons) revealed that the *rough* condition was rated significantly lower than

Table 4.1: Mean, median, and standard deviation of MUSHRA ratings by condition for all 37 participants.

Condition	Mean	Median	SD
Natural	33.6	28.0	27.6
Robotic	29.5	20.0	28.9
Rough	23.8	20.0	22.6

both natural ($W = 21.0$, $p < .001$) and robotic ($W = 140.5$, $p = .012$). The difference between natural and robotic was not significant ($W = 242.5$, $p = .235$). The condition effect therefore reflects primarily the suppression of the *rough* condition rather than a reliable perceptual distinction between the two lower-perturbation conditions.

This pattern was broadly consistent across individual vowels: Friedman tests run separately per vowel yielded significant effects for /ɔ/ ($\chi^2(2) = 13.54$, $p = .001$), /u/ ($\chi^2(2) = 9.52$, $p = .009$), /ɛ/ ($\chi^2(2) = 8.83$, $p = .012$), /e/ ($\chi^2(2) = 8.36$, $p = .015$), /a/ ($\chi^2(2) = 6.13$, $p = .047$), and /i/ ($\chi^2(2) = 6.28$, $p = .043$); the effect for /o/ did not reach significance ($\chi^2(2) = 5.69$, $p = .058$).

Vowel differences

Table 4.2 presents condition means by vowel. The back mid vowel /ɔ/ yielded the highest natural-condition mean ($M = 39.9$), followed by /a/ ($M = 37.6$), while /i/, /o/, and /u/ scored lowest across all conditions. The one departure from the ordering occurred for /a/, where the robotic condition ($M = 38.5$) marginally exceeded natural ($M = 37.6$). This difference is under one scale point and is consistent with the natural–robotic contrast reported above, considered to be not significant.

Low vowels (/a/, /ɔ/) consistently attracted higher ratings across all conditions than high vowels (/i/, /u/). This is consistent with the acoustic properties of the LF-source model. The first formant of low vowels is spectrally prominent and relatively independent of the high-frequency content that differentiates the synthesis

Table 4.2: Mean MUSHRA ratings by vowel and condition across all 37 participants.

Vowel	Natural	Robotic	Rough
/i/	28.6	26.0	21.9
/e/	35.7	27.5	24.8
/ɛ/	33.8	31.1	22.7
/a/	37.6	38.5	30.0
/ɔ/	39.9	34.8	26.2
/o/	29.0	26.0	21.4
/u/	30.7	22.9	19.8

conditions, whereas high vowels are more sensitive.

Group differences

By listener group (Table 4.3), listeners from Tuscany gave the highest overall means ($M = 36.5$, collapsed over vowels and conditions), followed by Todi ($M = 31.7$) and all other regions ($M = 24.7$). These differences are reported descriptively only. The Tuscany group comprised four participants, which is too few to support inferential analysis. The lower scores from all other regions may partly reflect lower engagement with the synthesis stimuli among participants from dialectal regions further from the central Italian dialect area.

Table 4.3: Mean MUSHRA ratings by listener group and synthesis condition.

Group	N	Natural	Robotic	Rough	Overall
Tuscany	4	38.6	37.7	33.2	36.5
Todi	16	37.1	31.5	26.6	31.7
Other	17	29.2	25.7	19.1	24.7

The ordering Tuscany ≥ Umbria > Other is consistent with the effect of regional familiarity; listeners within the central Italian dialect zone may find the synthesis less perceptually foreign (Clopper, 2021). Given the small sample size for the Tuscan group, its position cannot be reliably affirmed. Nonetheless, the prominence of the

Umbria group promotes the validity of the synthesis conditions used in Tasks 2 and 3.

Interpretation

There are two implications that arise from the results. First, the *rough* parameterization reliably reduces perceived naturalness relative to both alternatives. This is consistent with the known relationship between source perturbation and voice quality. Moderate aperiodicity contributes to a natural voice quality, whereas excessive perturbation is heard as roughness (D. H. Klatt and L. C. Klatt, 1990). Second, the natural and robotic conditions are not perceptually distinguishable ($p = .235$). The modest prosodic modeling applied in the *natural* condition does not yield a statistically reliable naturalness benefit over the plain LF baseline under these synthesis parameters and listener conditions.

These results nonetheless provide a sufficient basis for selecting the *natural* condition as the primary synthesis for Tasks 2 and 3. Although it is not significantly better than *robotic*, it is the highest-rated condition across all vowels and all listener groups, and there is no evidence that it is worse. Selecting it over *robotic* therefore carries no perceptual cost.

Methodological limitations

The Friedman test is interpretable as an ordinal comparison and is not affected by the absolute scaling of the ratings. However, a sampling-rate mismatch between the natural speech reference signal and the synthesized signals may have introduced a limitation to what the absolute scores can be taken to mean. The reference was drawn from speech sampled at 44.1 kHz, whereas all three synthesized conditions were generated at 10 kHz. Under ITU-R BS.1534, the visible reference defines the 100-point ceiling of the rating scale, and listeners calibrate their judgments against it (International Telecommunication Union, 2015). Because the reference contained spectral energy above 5 kHz that was absent from every test stimulus, no condition could approach 100 by design. The gap between any synthesized stimulus and the reference includes an irreducible component due to bandwidth rather than synthesis

quality. Absolute scores may not therefore be valid indicators of naturalness.

4.2 Task 2: Forced-choice identification

Task 2 consisted in a seven-alternative forced-choice vowel identification task to assess how accurately listeners mapped the formant-synthesized Todi vowels onto the seven-phoneme categories of the Italian vowel system. The stimuli were the seven synthesized Todi vowels, which presented three times each, yielding 21 trials per participant in randomized order. Participants selected their response from a screen displaying IPA labels of the seven vowels. The full sample of $N = 45$ participated; unlike Task 1, no non-engagement exclusion was applied to this task.

Results

Collapsed across vowels and groups, mean identification accuracy was 71.8% (chance = $1/7$ = 14.3%). Performance was, however, highly uneven across the individual vowels. Accuracy ranged from near-ceiling on /a/ (M = 99.1%) and /i/ (M = 89.5%) to near-chance on /ɔ/ (M = 34.2%). The distribution was clearly bimodal: four vowels over half of the vowels were identified at 72% or above, while the two open-mid vowels /ɛ/ and /ɔ/ performed substantially worse.

Table 4.4: Forced-choice identification accuracy (%) by vowel and listener group (N = 38; chance = 14.3%).

Vowel	Overall	Todi	Tuscany	Other
/a/	99.1	97.9	100.0	100.0
/i/	89.5	89.6	75.0	92.6
/e/	78.1	79.2	100.0	72.2
/u/	77.2	72.9	66.7	83.3
/o/	72.8	81.2	75.0	64.8
/ɛ/	51.8	64.6	58.3	38.9
/ɔ/	34.2	41.7	66.7	20.4
Overall	71.8	75.3	77.4	67.5

Table 4.4 shows per-vowel accuracy by listener group, and Table 4.5 presents the full confusion matrix for the seven vowels. The acoustic extremes (/i/ and /a/) of

the vowel space were identified with near-perfect accuracy. Moderate accuracy was observed for the high back vowel /u/ (77.2%), the closed-mid vowels /e/ (78.1%) and /o/ (72.8%), and for /ɛ/ (51.8%).

The two low-accuracy vowels showed distinct confusion profiles. For /ɛ/, 43.9% of responses were /e/, which for these vowels was the dominant error pattern; only 3.5% of responses were other vowels. For /ɔ/, the primary confusion was /o/ at 50.0%, with a secondary confusion of /a/ at 14.9%. This pattern suggests that for a subset of listeners the Todi /ɔ/ was perceived as having a degree of openness comparable to /a/ rather than /o/. No other vowel produced a dominant cross-category confusion. Secondary errors occurred at rates below 20% for the remaining five vowels.

Table 4.5: Confusion matrix for the identification task, showing the percentage of responses in each category by target vowel.

Target	/i/	/e/	/ɛ/	/a/	/ɔ/	/o/	/u/
/i/	**89.5**	7.9	0.9	0.9	0.0	0.9	0.0
/e/	0.0	**78.1**	16.7	0.9	0.9	1.8	1.8
/ɛ/	0.0	43.9	**51.8**	0.0	3.5	0.0	0.9
/a/	0.0	0.9	0.0	**99.1**	0.0	0.0	0.0
/ɔ/	0.0	0.0	0.9	14.9	**34.2**	50.0	0.0
/o/	0.9	0.0	0.9	2.6	18.4	**72.8**	4.4
/u/	0.9	0.9	0.0	0.0	6.1	14.9	**77.2**

Group differences

The three listener groups showed the clearest divergence on the two open-mid vowels. For /ɔ/, accuracy ranged from 66.7% in Tuscany to 41.7% in Todi to 20.4% in all other regions, a spread of 46 percentage points. For /ɛ/, the ordering was Todi (64.6%) > Tuscany (58.3%) > all other regions (38.9%). On the high-accuracy vowels (/a/, /e/), group differences were negligible. On /u/, the ordering was reversed (Other > Todi > Tuscany). Group differences are reported descriptively. The Tuscany group (N = 4) is too small for inferential analysis, and the all other regions group contains heterogeneous regional backgrounds.

Interpretation

The overall accuracy of 71.8% indicates successful identification of the synthesized Todi vowels at a rate well above chance. This suggests that the synthesis procedure and formant filters reliably reproduce vowel identity. The bimodal distribution of per-vowel accuracy (four vowels at 72% or above and two vowels below 55%) indicates that the mid-vowels are particularly difficult to identify. This result is to be expected, as /ɛ/ and /ɔ/ present the most acoustic variation across local Italian varieties (Canepari, 2026). The Todi realizations of these vowels also have the most pronounced departure from the ones measured by Ferrero et al. (1978).

In the present study, the synthesized stimuli are based on measured vowel productions from the Todi variety, which are acoustically more extreme than Standard Italian (SI) norms. The low identification accuracy for some vowels could be explained by the results of Johnson, Flemming, and Wright (1993): even these extreme Todi values might fall short of listeners' idealized, hyperarticulated targets. If a listener's internal target is more extreme than any natural production, stimuli synthesized at natural production norms will fail to reach that target, leaving the sound in an ambiguous perceptual zone between vowel categories. This helps explain why identification accuracy for these vowels remains low. Furthermore, static steady-state synthesis strips away the dynamic spectral changes that are crucial for accurate vowel identification (Hillenbrand et al., 1995). Therefore, it is plausible that for /ɛ/ and /ɔ/, the fixed formant values fall into an ambiguous acoustic region where missing dynamic cues are most desperately needed to resolve category membership.

The intermediate performance of the Todi listeners (41.7% for /ɔ/; 64.6% for /ɛ/) can also be understood as a result of a separate phenomenon. Unregistered dialect perception affects speakers who do not socially stereotype their own regional variety as a distinct dialect. This leads them to perceive their own speech as closer to a standard form than it actually is. Since, as Canepari (2026), all speakers in the central Italian dialect area are "native Italian speakers", is highly probable that the Todi speakers have internalized Standard Italian as their true phonetic target. This is supported by the relatively low ratings for daily dialect usage among the participants:

their own extreme acoustic pronunciations are not internalized as a regional accent with distinct phonetic realizations. As a result, their perceptual targets for /ɛ/ and /ɔ/ would be calibrated to Standard Italian. In summary, along this line of reasoning, even though the synthesized Todi vowels acoustically match the Umbrian speakers' actual real-world pronunciations, they may fall into perceptually ambiguous zones because the listeners are evaluating them against their internalized Standard Italian targets.

Methodological considerations

Two important methodological features of the identification task should be kept in mind. First, response options were presented as IPA symbols with carrier word references as examples. As Hillenbrand et al. (1995) notes, this leads to noticeable error rates for untrained participants. Furthermore, participants who are, as Canepari (2026) calls them, *dalfonici* (a pun of *daltonici*, "colorblind"), may not be able to distinguish the sounds anyways, let alone the interface itself. The extent to which this factor contributes to the observed confusions cannot be determined from the present data.

Second, the task used a single synthesis condition (the *natural* condition selected on the basis of Task 1 results) and static Todi formant targets. The identification accuracy reported here reflects perception of synthesized steady-state vowels at specific acoustic coordinates, not the identification of Todi dialect speech more broadly. Variability in natural Todi productions, such as vowels embedded in carrier words, would likely yield a different pattern of identification accuracy and confusions.

4.3 Task 3: Prototype selection

Task 3 presented each of the seven five-step synthesis continua to listeners and asked them to select the step that most closely matched their own habitual pronunciation of the target vowel, following the paradigm described in Section 3.9. The full sample of $N = 38$ completed the task. Before analysis, responses for /u/ were excluded on the grounds of a direction asymmetry that indicates a systematic position bias for that vowel. When the Standard Italian target was placed on the left of the response

panel (standard-left trials), the mean dialect score for /u/ was 2.37; when the Todi endpoint was placed on the left (dialect-left trials), the mean was 3.63, a difference of $\Delta = -1.26$. For all other vowels the corresponding asymmetry was $|\Delta| \leq 0.48$, indicating that the /u/ responses were dominated by a left-side preference rather than by genuine prototype selection. All results reported below are therefore based on the six remaining vowels.

Results

Collapsed across the six vowels and all listener groups, the mean per-participant dialect score was 3.57 ($SD = 0.51$), which is significantly above the continuum midpoint of 3.0 ($p < .001$). This indicates a mild but reliable overall preference for steps closer to the Todi endpoint of the continuum.

Table 4.6: Mean per-participant dialect score by listener group for six vowels (/u/ excluded). Dialect score ranges from 1 (most Standard Italian-like) to 5 (most Todi-like); the continuum midpoint is 3.

Group	*N*	Mean	*SD*
Todi	16	3.37	0.42
Tuscany	4	3.46	0.28
Other	18	3.78	0.56
Overall	38	3.57	0.51

A Kruskal-Wallis test on per-participant means revealed a significant group effect ($H(2) = 9.88$, $p = .007$). Comparisons with a U-test showed that the Other group selected more Todi-like steps than the Todi group ($U = 59.0$, $p = .003$). The Tuscany group did not differ significantly from either Other ($p = .070$) or Todi ($p = .70$). Although all three groups produced means above the continuum midpoint, the Todi and Tuscany groups hovered close to it (3.37 and 3.46, respectively), while the Other group showed a noticeably stronger preference for the Todi-side steps (3.78).

Vowel-level variation

Table 4.7 separates mean dialect scores by vowel and listener group. Two vowels stand out for large cross-group divergence. For /o/, the spread across groups was 1.89 scale points (Other: 4.39 vs. Tuscany: 2.50), and for /ɔ/ the ordering of the groups reversed relative to the overall trend. Here, the Other group selected the most Standard-Italian-like steps (M = 2.67), below the midpoint, while the Todi and Tuscany groups selected Todi-leaning steps (3.44 and 3.75, respectively). The vowels /a/ and, to a lesser extent, /ɛ/ showed more uniform cross-group distributions. For the front vowels /i/, /e/, and /ɛ/, the Other group consistently chose higher dialect scores than the Todi group.

Table 4.7: Mean dialect score by vowel and listener group (six vowels; /u/ excluded). Scores range from 1 (most SI-like) to 5 (most Todi-like).

Vowel	Todi	Tuscany	Other	Overall
/i/	3.19	3.25	4.00	3.58
/e/	3.69	3.50	4.39	4.00
/ɛ/	3.31	4.00	3.89	3.66
/a/	3.38	3.75	3.33	3.39
/ɔ/	3.44	3.75	2.67	3.11
/o/	3.19	2.50	4.39	3.68

Dialect usage

Within the Todi group, participants were stratified by self-reported dialect frequency. Those who reported never or rarely speaking dialect (questionnaire values 1–2; N = 7) had a mean dialect score of 3.31; those at the intermediate level (value 3; N = 5) scored 3.50; and those at high frequency (values 4–5; N = 4) scored 3.29. No systematic trend is apparent across the three levels, and the within-group variance is too large and the cell sizes too small to draw any inferential conclusions from this pattern. The absence of a gradient from low to high dialect frequency is itself consistent with the unenregistered variety interpretation developed in Section 4.3. That is, if Todi speakers do not internally register their variety as phonetically distinct from Standard Italian, frequency of dialect use may not be expected to

predict prototype location.

Sex

A sex difference was observed in the expected direction. A Mann–Whitney test ($U = 263.0$, $p = .015$) revealed that male participants ($N = 18$, $M = 3.77$) selected more Todi-like steps than female participants ($N = 20$, $M = 3.39$). However, this effect is affected by the group compositions. The Other group, which showed the highest dialect scores, comprises 13 male and 5 female participants, while the Todi group, which showed the lowest, comprises 4 male and 12 female participants. The sex difference is therefore reported descriptively as a potential confound.

Duration

Session duration was negatively correlated with mean dialect score in the expected direction (Spearman $\rho = -0.21$), which suggests that longer sessions may be associated with slightly more conservative, Standard-Italian-leaning selections; this correlation, however, did not reach significance ($p = .20$).

Interpretation

The main, and unexpected, finding of Task 3 is that the Todi group did not produce the highest dialect scores. If Todi listeners had strong perceptual prototypes calibrated to their own peripheral vowels they should have selected the most Todi-like continuum steps to describe their own pronunciation (Kuhl, 1991). Instead, the Other group, who have no direct experience with Todi speech, chose more Todi-like steps than the Todi group on average.

The framework of unenregistered dialect perception, as described by Clopper (2021), provides the most reasonable explanation for this trend. The present Todi data are consistent with this pattern. If Todi speakers have internalized Standard Italian formant targets as their mental norm for what a vowel "should" sound like, then they would select continuum steps closer to the SI endpoint as their prototypes, regardless of how their actual productions are distributed. Their mean of 3.37 lies only modestly above the midpoint, suggesting their prototypes do not strongly favor

either endpoint.

Accounting for the Other group is less straightforward. One possibility is that many non-Umbrian Italian varieties have vowel targets that are peripheral relative to Ferrero et al.'s (1978) Standard Italian measurements. The realization of mid vowels varies considerably across the peninsula (Canepari, 2026). If Other participants had regional prototypes for, say, /i/ or /e/ that were already more front or more open than the SI values, the Todi endpoint might have served as a better acoustic match for their own pronunciation than the SI endpoint. The per-vowel data provide partial support for this interpretation: the Other group selects strongly Todi-like steps for /i/, /e/, and /o/, but strongly SI-like steps for /ɔ/. This vowel-specific pattern suggests that the Other group's responses are driven by diverse prototype locations.

If listeners maintain phonetic prototypes that are more extreme than the SI averages, the most peripheral step on each continuum would consistently attract the highest rating (Johnson, Flemming, and Wright, 1993). However, this effect was originally demonstrated in a two-dimensional F_1–F_2 space, and the present continua are one-dimensional interpolations between two measured speaker-average values. It is not clear if using steps interpolated along the minimum Euclidean distance between the SI and Todi targets could have therefore biased the results. This account is therefore treated with caution.

Methodological considerations

Three methodological features of Task 3 bear on the interpretation of the results. First, as noted in Section 4.2, the response interface presented IPA symbols and carrier word anchors as category labels. The same limitation applies here: participants who cannot reliably map a symbol to a vowel category may not be selecting the step that matches their own pronunciation of that vowel but rather one that sounds most natural or familiar in isolation.

Second, the continua are one-dimensional linear interpolations in Bark space between two measured vowel averages. The acoustic difference between adjacent steps along a single formant dimension may not reflect the same perceptual distance as an

equivalent Euclidean distance in the full F_1–F_2 plane. In particular, the two targets represent average values across a small group of both Todi speakers and the reference values from Ferrero et al. (1978). Within-group variability in real speech likely spans a range broader than the interval between steps 1 and 5 on any continuum. Responses in the middle of the scale therefore do not unambiguously indicate an intermediate prototype, but rather the absence of a strong attractor anywhere in the sampled range.

Third, the exclusion of /u/ from the analysis reduces the vowels from seven to six. The /u/ exclusion was motivated by asymmetry that substantially exceeded the maximum observed for any other vowel. This asymmetry is consistent with a position-preference artifact rather than a systematic prototype judgment. Its inclusion would have distorted both per-participant averages and group comparisons. Further investigations into the asymmetry of /u/ is a question for future work.

Chapter 5

CONCLUSION

This thesis addressed two research questions concerning the tonic vowel system of the Todi variety of Italian. The measurements for this variety exhibit systematic deviations from the Standard Italian reference documented by Ferrero et al. (1978). The engineering question investigated the quality of the synthetic stimuli. The linguistic question asked whether listeners from Todi have different locations for their mental prototypes of the seven tonic vowels of Italian. The synthesis pipeline was presented in Chapter 2; the experimental design, including the Todi acoustic measurements, in Chapter 3; and the results and discussion in Chapter 4. This chapter summarizes the conclusions from each chapter, discusses their theoretical significance, identifies the main limitations of the approach, and proposes directions for future research.

5.1 Main findings

The parallel formant synthesizer developed for this study successfully implements the vowels in a computationally transparent and acoustically controllable form. The cascade-to-parallel gain calibration procedure aligns formant peak amplitudes to within 0.1 dB of the cascade reference across all seven Todi vowels. The alias-free frequency-domain LF source (Gobl, 2021) eliminates the aliasing distortion that is perceptually detectable at standard sampling rates (Wang and Gobl, 2023), ensuring that the glottal excitation introduces no spectral artifacts into the stimuli.

The MUSHRA naturalness evaluation confirmed the predicted ordering of the three synthesis conditions. The *rough* condition received significantly lower naturalness ratings than both the *natural* and *robotic* conditions ($p < .001$ and $p = .012$ respectively). The difference between *natural* and *robotic*, however, was not significant ($p = .235$). This pattern is consistent with the well-established relationship between source perturbation and perceived voice quality. Moderate aperiodicity is a property

of healthy phonation that contributes to naturalness, whereas excessive perturbation is perceived as roughness rather than as an enhancement (D. H. Klatt and L. C. Klatt, 1990).

The absolute rating levels were uniformly within the ITU-R *Poor* band; M = 33.6, 29.5, and 23.8 for natural, robotic, and rough respectively. As discussed in Section 4.1, this outcome is attributable to a bandwidth mismatch. The natural speech references were sampled at 44.1 kHz, whereas all synthesized stimuli were generated at 10 kHz. Because the reference defines the 100-point ceiling of the MUSHRA scale (International Telecommunication Union, 2015), no synthesized condition could approach the upper portion of the scale. The absolute scores are therefore not interpretable as direct measures of synthesis quality in isolation, however, the ordinal comparison alone is valid. On ordinal grounds, the *natural* condition is the highest-rated option across all vowels and all listener groups, as expected.

Vowel identification

The forced-choice identification task confirmed that the synthesized Todi vowels are categorized as Italian phonemes at a rate well above chance (71.8%; chance = 14.3%). This implies that the synthesis correctly reproduces the information needed for vowel identification. The corner vowels /a/ and /i/ were identified with near-ceiling accuracy (99.1% and 89.5% respectively), and moderate accuracy was obtained for /e/, /u/, and /o/ (78.1%, 77.2%, and 72.8%). The open-mid vowels /ɛ/ and /ɔ/ performed substantially below the remaining five, at 51.8% and 34.2%.

The dominant error pattern for both low-accuracy vowels was confusion with the adjacent closed-mid category: 43.9% of /ɛ/ responses were /e/, and 50.0% of /ɔ/ responses were /o/. The motivation for this result uncertain, although an important consideration is that the open/closed distinction in the mid-vowel pairs is the narrowest F_1 separation in the system (approximately 140–160 Hz from Ferrero et al. 1978). The absence of formant-transition cues is disadvantageous for the vowels whose category membership relies most heavily on them (Hillenbrand et al., 1995).

Listener group differences were most pronounced on the open-mid vowels. Central Italian listeners (Todi and Tuscany) outperformed those from other regions on both /ɛ/ and /ɔ/, consistent with greater perceptual sensitivity to the open/closed distinction in varieties that maintain it robustly. For the high-accuracy vowels, group differences were negligible.

Prototype selection

The prototype selection task yielded the most theoretically significant finding of the study. Against the prediction of the perceptual magnet effect (Kuhl, 1991), listeners from Todi did not select the most dialect-like continuum steps as their prototype. Their mean dialect score was 3.37 on a 1–5 scale from Standard Italian to Todi, only modestly above the continuum midpoint of 3.0. The group composed of all other regions except Tuscany produced a mean of 3.78, which is more Todi-like despite reporting having no experience with the Todi variety. A non-parametric group comparison confirmed a significant overall group effect and a significant pairwise difference between the Other and Todi groups.

The vowel-level analysis added nuance to this pattern. The Other group selected strongly Todi-like steps for the front vowels /i/ (M = 4.00), /e/ (M = 4.39), and /ɛ/ (M = 3.89), but markedly SI-like steps for /ɔ/ (M = 2.67, below the midpoint). The large cross-group spread for /o/ (1.89 scale points) and the reversal for /ɔ/ suggest that the Other group's responses are driven by diverse regional prototype locations.

Within the Todi group, no systematic gradient from low to high dialect frequency was observed. This is consistent with the proposal that self-reported dialect use does not predict prototype location when the variety is not socially registered as phonetically distinct from Standard Italian.

5.2 Limitations

The interpretation of the results is limited by several methodological constraints. The Todi acoustic data derive from only eight speakers, which sufficient for a preliminary description but inadequate for fully characterizing the variety. Formant values were

manually extracted by a single annotator without reported inter-rater reliability. The Standard Italian reference (Ferrero et al., 1978) lacks sex-disaggregated data, and the assumption of matched speaker groups remains untestable.

The synthesis employed static steady-state vowels with fixed formant frequencies at a 250 ms duration. Omitting dynamic formant transitions known to be crucial for vowel identity (Hillenbrand et al., 1995). This under-represents natural acoustic cues. The 10 kHz sampling rate further limits the spectrum to 5 kHz, sufficient for identity but creating a quality gap versus 44.1 kHz natural speech in MUSHRA evaluations.

Listener groups were unbalanced in size and sex composition, with the small Tuscany group ($n = 4$) excluded from comparisons. The heterogeneous "Other" group spans twelve regions, complicating interpretation. IPA-based response options may have challenged untrained participants, inflating confusions. The /u/ vowel was excluded from prototype analysis due to extreme directional asymmetry.

5.3 Future research

Future research should include larger, sex-balanced speaker and listener samples. Bigger sample sizes would enable sex-, age-, and city/regional-specific analyses. Extending data collection to other Umbrian varieties (e.g., Foligno, Spoleto) would allow comparison of measurements to see if the Todi speech data is localized or part of a larger Umbrian or central Italian phenomenon.

The overall conclusion is that the data are inconclusive but point to there not being a clear perceptual difference between Standard Italian and the variety of Italian spoken in Todi, or to the variety being socially unenregistered. More research into the social conditions under which this dialect is perceived is therefore considered a fundamental next step.

BIBLIOGRAPHY

Acoustical Society of America (2018). *Ethical Principles of the Acoustical Society of America for Research Involving Human and Non-Human Animals in Research and Publishing and Presentations*. Acoustical Society of America. URL: https://acousticalsociety.org/ethical-principles/ (visited on 06/08/2026).

Bertinetto, Pier Marco and Michele Loporcaro (2005). "The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome". In: *Journal of the International Phonetic Association* 35.2, pp. 131–151. DOI: 10.1017/S0025100305002148.

Boersma, Paul and David Weenink (2024). *Praat: doing phonetics by computer*. Computer program. https://www.praat.org/. Version 6.4.06.

Canepari, Luciano (2026). *Dizionario di pronuncia italiana neutra (DÍPiN)*. Venice: Università di Venezia.

Clopper, Cynthia G. (2021). "Perception of dialect variation". In: *The Handbook of Speech Perception*. Ed. by Jennifer S. Pardo et al. 2nd ed. Hoboken, NJ: Wiley-Blackwell, pp. 335–365.

Cooper, Franklin S. et al. (1952). "Some experiments on the perception of synthetic speech sounds". In: *Journal of the Acoustical Society of America* 24.6, pp. 597–606. DOI: 10.1121/1.1906940.

Delattre, Pierre et al. (1952). "An Experimental Study of the Acoustic Determinants of Vowel Color; Observations on One- and Two-Formant Vowels Synthesized from Spectrographic Patterns". In: *WORD* 8, pp. 195–210. URL: https://api.semanticscholar.org/CorpusID:147079156.

Fant, Gunnar (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

– (1995). "The LF-model revisited. Transformations and frequency domain analysis". In: *Speech, Transmission Laboratory Quarterly Progress and Status Report* 36.2–3, pp. 119–156.

Fant, Gunnar, Johan Liljencrants, and Qi-guang Lin (1985). "A four-parameter model of glottal flow". In: *Speech, Transmission Laboratory Quarterly Progress and Status Report* 26.4, pp. 1–13.

Farrús, Mireia, Javier Hernando, and Pascual Ejarque (2007). "Jitter and shimmer measurements for speaker recognition". In: *Proc. Interspeech 2007*. Antwerp, Belgium, pp. 778–781. DOI: 10.21437/Interspeech.2007-147.

Ferrero, Franco E. et al. (1978). "Some acoustic characteristics of the Italian vowels". In: *Italian Linguistics* 4, pp. 87–95.

Flanagan, James L. (1957). "Estimates of the maximum precision necessary in quantizing certain "dimensions" of vowel sounds". In: *Journal of the Acoustical Society of America* 29.4, pp. 533–534. DOI: 10.1121/1.1908957.

Gobl, Christer (2017). "Reshaping the transformed LF model: generating the glottal source from the waveshape parameter R_d". In: *Proc. Interspeech 2017*. Stockholm, Sweden, pp. 3008–3012. DOI: 10.21437/Interspeech.2017-1140.

– (2021). "The LF model in the frequency domain for glottal airflow modelling without aliasing distortion". In: *Proc. Interspeech 2021*. Brno, Czechia, pp. 2651–2655. DOI: 10.21437/Interspeech.2021-1625.

Goudbeek, Martijn et al. (2005). "Acquiring Auditory and Phonetic Categories". In: *Handbook of Categorization in Cognitive Science*. Ed. by Henri Cohen and Claire Lefebvre. Amsterdam: Elsevier, pp. 497–513. ISBN: 9780080446127. DOI: 10.1016/B978-008044612-7/50077-9.

Hillenbrand, James et al. (1995). "Acoustic characteristics of American English vowels". In: *Journal of the Acoustical Society of America* 97.5, pp. 3099–3111. DOI: 10.1121/1.411872.

Holmes, John N. (1983). "Formant synthesisers: cascade or parallel?" In: *Speech Communication* 2.4, pp. 251–273. DOI: 10.1016/0167-6393(83)90018-8.

International Telecommunication Union (2015). *Method for the subjective assessment of intermediate quality level of audio systems*. Recommendation BS.1534-3. ITU-R.

Johnson, Keith, Edward Flemming, and Richard Wright (1993). "The hyperspace effect: phonetic targets are hyperarticulated". In: *Language* 69.3, pp. 505–528. DOI: 10.2307/416697.

Johnson, Keith and Matthias J. Sjerps (2021). "Speaker Normalization in Speech Perception". In: *The Handbook of Speech Perception*. Ed. by Jennifer S. Pardo et al. 2nd ed. Hoboken, NJ: John Wiley & Sons, pp. 145–176. DOI: 10.1002/9781119184096.ch6.

Klatt, Dennis H. (1980). “Software for a cascade/parallel formant synthesizer”. In: *Journal of the Acoustical Society of America* 67.3, pp. 971–995. DOI: 10.1121/1.383940.

Klatt, Dennis H. and Laura C. Klatt (1990). “Analysis, synthesis, and perception of voice quality variations among female and male talkers”. In: *Journal of the Acoustical Society of America* 87.2, pp. 820–857. DOI: 10.1121/1.398894.

Kuhl, Patricia K. (1991). “Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not”. In: *Perception & Psychophysics* 50.2, pp. 93–107. DOI: 10.3758/BF03212211.

Lange, Kristian, Simone Kühn, and Elisa Filevich (June 2015). “"Just Another Tool for Online Studies” (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies”. In: *PLOS ONE* 10.6, pp. 1–14. DOI: 10.1371/journal.pone.0130834. URL: https://doi.org/10.1371/journal.pone.0130834.

Leeuw, Joshua R. de, R. A. Gilbert, and B. Luchterhandt (2023). “jsPsych: Enabling an Open-Source Collaborative Ecosystem of Behavioral Experiments”. In: *Journal of Open Source Software* 8.85, p. 5351. DOI: 10.21105/joss.05351. URL: https://joss.theoj.org/papers/10.21105/joss.05351.

Moretti, Nanni (1987). *11: Umbria / di Giovanni Moretti*. ita. Consiglio nazionale delle ricerche, Centro di studio per la dialettologia italiana. Pisa: Pacini.

Oord, Aäron van den et al. (2016). “WaveNet: A Generative Model for Raw Audio”. In: *Proc. 9th ISCA Speech Synthesis Workshop (SSW9)*. Sunnyvale, CA. DOI: 10.21437/SSW.2016-11.

Oppenheim, Alan V., Alan S. Willsky, and S. Hamid Nawab (1996). *Signals & systems (2nd ed.)* USA: Prentice-Hall, Inc. ISBN: 0138147574.

Perrotin, Olivier et al. (2023). “The Blizzard Challenge 2023”. In: *18th Blizzard Challenge Workshop*, pp. 1–27. DOI: 10.21437/Blizzard.2023-1.

Peterson, Gordon E. and Harold L. Barney (1952). “Control methods used in a study of the vowels”. In: *Journal of the Acoustical Society of America* 24.2, pp. 175–184. DOI: 10.1121/1.1906875.

Pisoni, David B. (1973). “Auditory and phonetic memory codes in the discrimination of consonants and vowels”. In: *Perception & Psychophysics* 13.2, pp. 253–260. DOI: 10.3758/BF03214136.

Raphael, Lawrence J., Gloria J. Borden, and Katherine S. Harris (2011). *Speech Science Primer: Physiology, Acoustics, and Perception of Speech*. 6th ed. Philadelphia: Wolters Kluwer / Lippincott Williams & Wilkins.

Rosner, B. S. and J. B. Pickering (1994). *Vowel Perception and Production*. Oxford and New York: Oxford University Press. ISBN: 9780198521385.

Schoentgen, Jean (2001). "Stochastic models of jitter". In: *Journal of the Acoustical Society of America* 109.4, pp. 1631–1650. DOI: 10.1121/1.1350557.

– (2003). "Decomposition of vocal cycle length perturbations into vocal jitter and vocal microtremor, and comparison of their size in normophonic speakers". In: *Journal of Voice* 17.2, pp. 114–125. DOI: 10.1016/S0892-1997(03)00014-6.

Schoentgen, Jean and Philipp Aichinger (2019). "Analysis and synthesis of vocal flutter and vocal jitter". In: *Proc. Interspeech 2019*. Graz, Austria, pp. 2518–2522. DOI: 10.21437/Interspeech.2019-1998.

Taylor, Paul (2009). *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511816338.

Traunmüller, Hartmut (1990). "Analytical expressions for the tonotopic sensory scale". In: *Journal of the Acoustical Society of America* 88.1, pp. 97–100. DOI: 10.1121/1.399849.

Ugoccioni, Nicoletta and Marcello Rinaldi (2001). *Vocabolario del dialetto di Todi e del suo territorio*. Vol. 12. Opera del Vocabolario Dialettale Umbro. Todi: Amministrazione Comunale di Todi. 732 pp.

Varadhan, Praveen Srinivasa et al. (2025). "Rethinking MUSHRA: Addressing Modern Challenges in Text-to-Speech Evaluation". In: arXiv: 2411.12719 [cs.CL]. URL: https://arxiv.org/abs/2411.12719.

Vignuzzi, Ugo (1997). "Lazio, Umbria and the Marche". In: *The Dialects of Italy*. Ed. by Martin Maiden and Mair Parry. London: Routledge. Chap. 37, pp. 311–320.

Wang, Zihan and Christer Gobl (2023). "The perceptual effects of aliasing distortion in glottal flow modelling". In: *Proc. 20th International Congress of Phonetic Sciences (ICPhS)*. Prague, Czechia, pp. 1801–1805.

Appendix A

CODE LISTINGS

```
import numpy as np
from numpy.polynomial.polynomial import polyadd

class TF:

    def __init__(self, b, a):
        self.b = np.asarray(b, dtype=float)
        self.a = np.asarray(a, dtype=float)

    def __mul__(self, other):
        if isinstance(other, TF):
            return TF(np.convolve(self.b, other.b),
                      np.convolve(self.a, other.a))
        return TF(self.b * other, self.a)

    def __rmul__(self, other):
        return self * other

    def __imul__(self, other):
        total = self * other
        self.b = total.b
        self.a = total.a
        return self

    def __add__(self, other):
        b = polyadd(np.convolve(self.b, other.a),
```

```
                    np.convolve(other.b, self.a))
        a = np.convolve(self.a, other.a)
        return TF(b, a)

    def __iadd__(self, other):
        total = self + other
        self.b = total.b
        self.a = total.a
        return self

    def __repr__(self):
        return f"TF(b={self.b}, a={self.a})"
```

Listing A.1: The Transfer Function (TF) class with overloaded operators for series and parallel interconnection of LTI systems.

```
def resonator(F, B, Ts):

    R = np.exp(-np.pi * B * Ts)
    theta = 2 * np.pi * F * Ts

    B_coeff = 2 * R * np.cos(theta)
    C_coeff = -(R**2)
    A_coeff = 1 - B_coeff - C_coeff

    return TF([A_coeff], [1.0, -B_coeff, -C_coeff])
```

Listing A.2: Python function for generating the transfer function of a second-order digital resonator using the TF class.

```
def pre_emphasis(alpha=0.9):
    return TF([1.0, -alpha], [1.0])
```

```
def differentiator():
    return TF([1.0, -1.0], [1.0])
```

Listing A.3: Convenience functions for common filters using the TF class (pre-emphasis and differentiator).

```
def cascade(F, B, Ts, n_res=5):

    H = TF([1], [1])

    for i in range(n_res):

        H_res = resonator(F[i], B[i], Ts)

        H *= H_res

    return H

def parallel(F, B, Ts, G=None, n_res=3):

    if G is None:
        G = np.ones(n_res)

    H = TF([0], [1])

    H_filter = pre_emphasis()

    for i in range(n_res):

        H_res = resonator(F[i], B[i], Ts)

        if i > 0:
```

```
            H_filter = differentiator()

        polarity = (-1) ** i

        H += polarity * G[i] * H_filter * H_res

    return H
```

Listing A.4: Cascade and parallel filters implemented with the TF class.

INDEX