Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1689–1694
Marseille, 11–16 May 2020

How to Compare Automatically Two Phonological Strings: Application to Intelligibility Measurement in the Case of Atypical Speech

Ghio[1] A., Lalain[1] M., Giusti[1] L., Fredouille[2] C., Woisard[3] V.

(1) Aix-Marseille Univ, LPL, CNRS, Aix-en-Provence, France
(2) LIA, Avignon University, France
(3) UT2J, Octogone-Lordat, Toulouse University & Toulouse Hospital, France

alain.ghio@lpl-aix.fr, muriel.lalain@lpl-aix.fr, corinne.fredouille@univ-avignon.fr, woisard.v@chu-toulouse.fr

Abstract

Atypical speech productions, regardless of their origins (accents, learning, pathology), need to be assessed with regard to "typical" or "expected" productions. Evaluation is necessarily based on comparisons between linguistic forms produced and linguistic forms expected. In the field of speech disorders, the intelligibility of a patient is evaluated in order to measure the functional impact of his/her pathology on his/her oral communication. The usual method is to transcribe orthographic linguistic forms perceived and to assign a global and imprecise rating based on their correctness or incorrect. To obtain a more precise evaluation of the production deviations, we propose a measurement method based on phonological transcriptions. An algorithm computes automatically and finely the distances between the phonological forms produced and expected from cost matrices based on the differences of features between phonemes. A first test of this method among a large population of healthy speakers and patients treated for cancer of the oral and pharyngeal cavities has proved its validity.

Keywords: intelligibility, phonological features, clinical phonetics

1. Introduction

We call "atypical speech" oral utterances deviating from a regular form in their pronunciation. If the notion of standard pronunciation remains a questionable notion when it touches on regional or sociologically marked forms of speech, it remains accepted in the case of learners of a foreign language (Kang et al., 2018) or in the case of motor-speech disorders or organic speech sound disorders.

In functional assessment of patients with speech disorder, intelligibility is a key parameter in, for example, dysarthria (Kent, 1992), head and neck cancer (Meyer et al., 2004) or speech production after cochlear implantation. Several methods of speech perception assessment are available to measure the severity of speech production disorders. Kent (1992) defines intelligibility as "the *degree* to which the speaker's intended *message* is recovered by the *listener"*. This author defines also as "item identification" the perceptual objective measurement, which is usually the percentage of items that are accurately recognized by a listener. In such a context, the atypical speaker can be asked to read a list of words or phrases, and the examiner writes down what (s)he has understood; the transcription is compared against the original list, and a score is calculated as the percentage of correctly understood items.

The transcription of the production as well as the target form are generally available in an orthographic form. However, if we are interested primarily in oral production and as this oral production is sometimes restricted to isolated words, it is ultimately more important to place the analysis at the phonological level. For instance, if the target word is "chaine" ("string") and the oral production is transcribed as "chêne" ("oak"), the intelligibility must be considered as perfect because these two words in French are produced in the same way /ʃɛnə/. Similarly, if the target word is "poule" ("hen") but the listener perceives "boule ("ball"), the error is less important than if the item was transcribed "brosse" ("brush") because in the first case, there is only a minimal error on the first phoneme while there are many differences in the second case. It seems important to go beyond the simple algorithm that provides 0 if the items are identical or 1 if there is a difference.

2. The algorithm to compare two phonological strings

Alignment of phonological sequences presupposes transcription of speech into discrete phonemic units and differs from matching of utterances in speech recognition. The alignment algorithm needs two components: a metric for measuring distance between phonemes and a process to find the best alignment (Kondrak, 2003). To do so, we used a Wagner-Fischer algorithm (Wagner & Fischer, 1974) that integrates the phenomena of insertion, elision, and unit substitution (Figure 1).

In our case, the calculation of Levenshtein distance bears on phonemes rather than orthographic forms, as it seemed important to us to establish a local distance between units (Ghio et al., 1995). Indeed, for the orthographic forms, traditionally, the distance between 2 graphemes is 0 if they are equal and 1 if they are different. In the case of phonemes, it is possible to be more specific; for example, the confusion between 2 vowels does not have the same weight in terms of error of production as that between a vowel and a voiceless consonant.

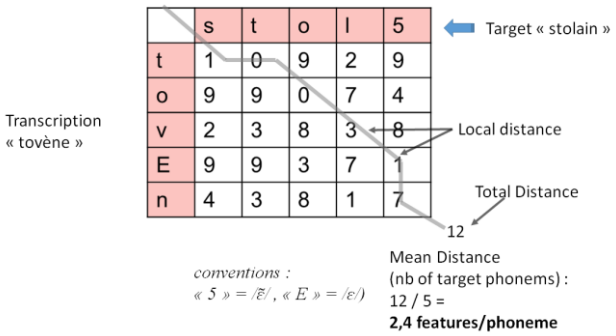

Figure 1 : Comparison of 2 phonological strings by the Wagner-Fischer algorithm (Conventions: « 5 » = /ɛ̃/, « E » = /ɛ/)

3. The cost matrix between phonemes

3.1 Metric

The comparison phase between 2 phonological strings requires the use of an inter-phoneme cost matrix which aims at defining the confusion between /p/ and /b/ (only one feature deviation), phonetically less important in terms of feature deviations between /f/ and /l/ (3 features).

The "cost" matrix is thus a table that contains the degree of dissimilarity between phonemes. It contains, for French, the 35 phonemes /a i u o ɔ e ɛ y œ ø ɔ̃ ɑ̃ ɛ̃ œ̃ p t k b d g f s ʃ v z ʒ m n l R j w ɥ ɲ ŋ/ to which are added various archiphonemes: Ô = /o/ or /ɔ/, Ê=/e/or/ɛ/, Û=/ø/ or /œ/, µ= /ɛ̃/ or /œ̃/, & = /e/ or /ɛ/ or /ø/ or /œ/. For the coding of phonological units in computer format, we used the convention of lexique.org (New et al., 2001) which allows the coding of a unit on one character, contrary to the SAMPA coding on which the correspondence is done on 1 or 2 characters.

To form the matrix, two strategies can be adopted:

- A measurement based on data. In this case, automatic procedures statistically calculate the average difference between phonemes. It is then a question of choosing a representative corpus as well as a relevant method of comparison.
- A measure based on knowledge. In this case, the distance between phonemes is attributed a priori from known data.

As the results of our intelligibility tests can be used as a basis for learning measurements resulting from automatic processing, we wanted to avoid a form of circularity and therefore discarded the first solution to favor the second method.

In order to reduce its arbitrary aspect, we have based the comparison on Feature theory (Jakobson et al., 1951).

According to this theory, phonemes can be characterized into a set of features that distinguishes them. It is easy to construct, from this decomposition, a multidimensional space in which each phoneme is geometrically located. The notion of feature imposing a binary status (present or absent), the coordinates of the phonemes in the multidimensional space only take values 0 or 1. In order to measure a distance in this multidimensional space, we decided to use a Manhattan distance, which is very simple because it consists in counting the number of different features between two phonemes.

There is two methods to characterize the phonemes with features :

- A single space with the same abstract features for both vowels and consonants (Dutrey et al, 2016)
- Separate spaces for vowels and consonants (Ghio, 1997) with features more phonetically and acoustically based.

With the first solution (Dutrey et al, 2016), we can obtain an unexpected short distance (d = 2) for instance between /e/ (continuous, coronal, vocalic, voiced) and /z/ (consonant, continuous, coronal, voiced) while the distance is higher (d=4) between /e/ and /u/ (continuous, dorsal, vocalic, voiced, high, rounded) which is not phonetically and acoustically pertinent. We preferred the phonetically-based proposition of (Ghio, 1997).

3.2 Cost matrix for vowels

Figure 2 and Table 1 present the characteristics of French vowels into features according to Chomsky and Halle (1968). We replaced the Chomsky feature [+/- low] with [+/- open] because less easily confused with the feature [+/- high], which is not the opposite of the feature [+/- low]. In this context, the mid vowels / e ø o / are [-high; -low] and oppose respectively / ɛ œ ɔ / which are [+ low], that is to say [+ open] in our denomination. The tree decomposition (Figure 2) makes it possible to highlight the notion of archiphoneme, that is to say the under-specification of a feature. Thus, the archiphonemes Ê = {e, ɛ}, Û = {œ, ø}, Ô = {o, ɔ} are units whose aperture feature is not specified; similarly, µ = {œ, ɛ} and & = {Ê, Û} neutralize the labialization feature. This characterisation thus makes it possible to draw up a matrix of distances between vowels (Table 2).

	a	i	u	o	e	y	ø	ɛ	ɔ	œ	Ô	Û	Ê	&	ɑ̃	ɛ̃	ɔ̃	œ̃	µ
nasal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
back	1	0	1	1	0	0	0	0	1	0	1	0	0	0	1	0	1	0	0
high	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
round	0	0	1	1	0	1	1	0	1	1	1	1	0		0	0	1	1	
open	1	0	0	0	0	0	0	1	1	1					1	1	1	1	1

Table 1 : phonetic features of French vowels (matrix form)

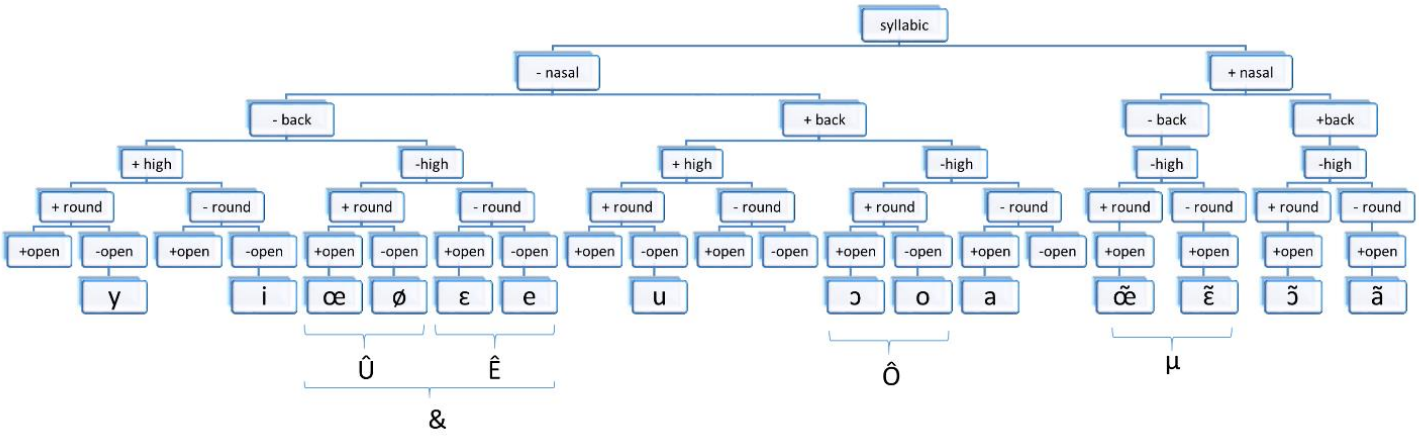

Figure 2 : phonetic features of French vowels (tree form)

	a	i	u	o	e	y	ø	ɛ	ɔ	œ	ã	ɛ̃	ɔ̃	œ̃	Ê	Ô	Û	µ	&
a	0	3	3	2	2	4	3	1	1	2	1	2	2	3	1	1	2	2	1
i	3	0	2	3	1	1	2	2	4	3	4	3	5	4	1	3	2	3	1
u	3	2	0	1	3	1	2	4	2	3	4	5	3	4	3	1	2	4	2
o	2	3	1	0	2	2	1	3	1	2	3	4	2	3	2	0	1	3	1
e	2	1	3	2	0	2	1	1	3	2	3	2	4	3	0	2	1	2	0
y	4	1	1	2	2	0	1	3	3	2	5	4	4	3	2	2	1	3	1
ø	3	2	2	1	1	1	0	2	2	1	4	3	3	2	1	1	0	2	0
ɛ	1	2	4	3	1	3	2	0	2	1	2	1	3	2	0	2	1	1	0
ɔ	1	4	2	1	3	3	2	2	0	1	2	3	1	2	2	0	1	2	1
œ	2	3	3	2	2	2	1	1	1	0	3	2	2	1	1	1	0	1	0
ã	1	4	4	3	3	5	4	2	2	3	0	1	1	2	2	2	3	1	2
ɛ̃	2	3	5	4	2	4	3	1	3	2	1	0	2	1	1	3	2	0	1
ɔ̃	2	5	3	2	4	4	3	3	1	2	1	2	0	1	3	1	2	1	2
œ̃	3	4	4	3	3	3	2	2	2	1	2	1	1	0	2	2	1	0	1
Ê	1	1	3	2	0	2	1	0	2	1	2	1	3	2	0	2	1	1	0
Ô	1	3	1	0	2	2	1	2	0	1	2	3	1	2	2	0	1	2	1
Û	2	2	2	1	1	1	0	1	1	0	3	2	2	1	1	1	0	1	0
µ	2	3	4	3	2	3	2	1	2	1	1	0	1	0	1	2	1	0	1
&	1	1	2	1	0	1	0	0	1	0	2	1	2	1	0	1	0	1	0

Table 2 : Vowel cost matrix (⇔ number of different features between vowels)

3.3 Cost matrix for consonants

In the characterization of French consonants, a number of features are clearly defined:

- The sonorant feature (+/- sonant) distinguishes the obstruents (occlusives and fricatives: -sonant) from the liquid consonants {l, R}, nasal {m n ɲ} and semi-vowels {j w ɥ}: + sonant
- The nasal feature distinguishes the nasal (+ nasal) consonants from the oral (-nasal)
- The voice feature distinguishes voiced (voiced) consonants from voiceless ones
- The continuant feature distinguishes occlusives (-cont) from fricatives (+ cont).

Among vowel consonants, Chomsky and Halle (1968) state p.317 that nasal occlusives are considered interrupted (-cont). The authors finally point out that the case of /l/ and /r/ is complex but ends up proposing a feature (+ cont) to /r/ and (-cont) to /l/. This characterization is confirmed in Clements (2005, p.47).

On the other hand, the features relating to the place of articulation of the consonant raise several issues. Indeed, according to the International Phonetic Alphabet (www.internationalphoneticalphabet.org), the French consonants are articulated according to 7 different places of articulation which can be grouped into 3 broad categories: the labials, the dentals and the velopalatals (Table 3).

	Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Palatal	Velar
Plosive	p b			t d			k g
Nasal	m			n		ɲ	ŋ
Fricative		f v		s z	ʃ ʒ		
approximant				l		j	

Labials Dentals Velopalatals

Table 3 : place of articulation of the French consonants (IPA)

In a phonological-based approach, Chomsky and Halle (1968) propose p.223 a characterization according to the two +/- coronal (tongue tip) and +/- anterior features, which gives Table 4 below. In this characterization, /p/ (+ coronal, + anterior) differs from /t/ (+ coronal) by one feature and similarly from /k/ (-anterior). On the other hand, /t/ (+ coronal, + anterior) differs from /k/ (-coronal, - anterior) by two features, which is not very satisfactory from an articulatory point of view where it would seem logical to respect the order /ptk/, that is to say /t/ equidistant from /p/ and /k/, /p/ and /k/ being more distant.

	+ coronal	-coronal
+anterior	Dental : t d s z	Labial : p b f v
-anterior	Palato-alveolar : neither specimen in French	Velar : k g (ʃ ʒ)

Table 4 : place of articulation features (Chomsky et Halle,1968)

Clements (2005) proposes a characterization into 3 exclusive features: labial, coronal and dorsal which directly reflects the 3 places described in Table 3. We estimate that there is overspecification because 2 features are sufficient to code 3 states. We finally opted for the work of Jakobson et al. (1951) which proposes two acoustic features permitting an adequate distinction:

- The compact / diffuse feature: "the consonant articulated against the hard or soft palate" (Jakobson et al., 1951, p.27)

- The grave / acute feature: "gravity characterizes labial consonants as against dentals, plus velars vs. palatals" (Jakobson et al., 1951, p. 30)

We finally obtained the characteristics of consonants into features (Table 5) and the matrix of distances between consonants (Table 6).

	p	t	k	b	d	g	f	s	ʃ	v	z	ʒ	m	n	ɲ	l	R	j	w	ɥ
sonorant (vocalic)	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
continuant	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	1	1	1	1
nasal	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
voiced	0	0	0	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1
compact	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0
acute	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	1	1	1	0	0

Table 5 : phonological features of French consonants

	p	t	k	b	d	g	f	s	ʃ	v	z	ʒ	m	n	ɲ	l	R
p	0	1	2	1	2	3	1	2	3	2	3	4	3	4	5	3	4
t	1	0	1	2	1	2	2	1	2	3	2	3	4	3	4	2	3
k	2	1	0	3	2	1	3	2	1	4	3	2	5	4	3	3	4
b	1	2	3	0	1	2	2	3	4	1	2	3	2	3	4	2	3
d	2	1	2	1	0	1	3	2	3	2	1	2	3	2	3	1	2
g	3	2	1	2	1	0	4	3	2	3	2	1	4	3	2	2	3
f	1	2	3	2	3	4	0	1	2	1	2	3	4	5	6	4	3
s	2	1	2	3	2	3	1	0	1	2	1	2	5	4	5	3	2
ʃ	3	2	1	4	3	2	2	1	0	3	2	1	6	5	4	4	3
v	2	3	4	1	2	3	1	2	3	0	1	2	3	4	5	3	2
z	3	2	3	2	1	2	2	1	2	1	0	1	4	3	4	2	1
ʒ	4	3	2	3	2	1	3	2	1	2	1	0	5	4	3	3	2
m	3	4	5	2	3	4	4	5	6	3	4	5	0	1	2	2	3
n	4	3	4	3	2	3	5	4	5	4	3	4	1	0	1	1	2
ɲ	5	4	3	4	3	2	6	5	4	5	4	3	2	1	0	2	3
l	3	2	3	2	1	2	4	3	4	3	2	3	2	1	2	0	1
R	4	3	4	3	2	3	3	2	3	2	1	2	3	2	3	1	0

Table 6 : Consonant cost matrix (⇔ number of different features between consonants)

The semi-consonants / j w ɥ / have been placed identically to their equivalent / i u y / less syllabic feature (-syll). Indeed, these phonemes alone cannot constitute a syllable (Chomsky & Halle, 1968). In their distances to the consonants, they have been characterized as presented in Table 5.

In relation to the vowels, the consonants have been placed at a distance greater than the maximum distance between vowels (d = 6). Taking into account the classification of Dell (1985), we then respected the following hierarchy:

Vowels <Liquid <Nasal < voiced Obstruents< unvoiced

<table>
<tr><td rowspan="3">non-syllabic</td><td rowspan="3">consonantic</td><td rowspan="2">non-vocalic</td><td>unvoiced</td><td>Unvoiced obstruents</td></tr>
<tr><td rowspan="3">voiced</td><td>Voiced obstruents</td></tr>
<tr><td rowspan="2">vocalic</td><td>Liquid and nasal consonants</td></tr>
<tr><td>non-consonantic</td><td></td><td>semi-vowels</td></tr>
<tr><td>syllabic</td><td></td><td></td><td></td><td>vowels</td></tr>
</table>

Table 7 : macro categories of consonants (Dell, 1985)

Finally, we obtained a "cost" matrix which contains the degree of dissimilarity, in number of features, between the 35 phonemes selected for French.

4. Application to intelligibility measurement

The previously described method was used in the context of the Carcinologic Speech Severity Index (C2SI) project, which aims to obtain a measure of the impact of oral and pharyngeal cancer treatments on speech production (Astesano et al, 2018). Indeed, in speech disorders, the phonetic realization of linguistic units is often different from the expected forms that occur in a normal speech. The degree of difference between the production and the expected form is an important issue for assessing the severity of the disorder and the success or failure of the communication (Connolly, 1997).

We call *Perceived-Phonological Deviation* (PPD) the distance between the expected sequence and the transcribed one. In our case, this PPD score is equal to the average number of phonological features misidentified by the listeners due to the articulatory disorders of the speakers.

In the C2SI project, we selected 126 speakers (41 healthy subjects and 85 patients) recorded in the Oncopole Hospital in Toulouse. Each speaker produced 52 random pseudo-words from a list of 89346 possible forms. 40 listeners transcribed these productions. The orthographic transcriptions were phonetized with the LIAPHON algorithm (Bechet, 2001) and compared to the expected phonetic forms of the pseudo-words by the algorithm described previously. The overall results (Table 8) show that the forms perceived in healthy subjects are on average at a distance of 0.48 non-identified features per phoneme compared to the expected forms whereas this distance rises to 1.28 for the patients. The difference is significant (ANOVA; $p < 0.01$). The results obtained on healthy subjects show that acoustico-phonetic decoding without lexical access is not perfect even on "normal" speech. By plotting a sensitivity / specificity curve ("ROC curve"), we explored the measure of the performance of a binary classifier that would distinguish normal / dysfunction based on the PPD score. The area under the ROC curve (AUC) which allows appreciating the quality of the classifier is 0.94, which corresponds to a high precision. This test therefore seems discriminating as regards the measurement of articulatory performance of speakers.

	N	Mean PPD	Standard dev
Healthy	41	0.48	0.22
Patients	85	1.28	0.63

Table 8 : Perceived-Phonological Deviation score between healthy speakers and patients (the unit of PPD score is in number of misidentified feature per phoneme)

5. Conclusion

We propose a method to compare automatically two phonological strings. The metric for measuring distance between phonemes is based on characterization of units by phonological features. We distinguished the spaces for vowels and consonants. The distance is a simple Manhattan distance where each feature has the same weight. The process to find the best alignment is a data time warping algorithm.

This method seems to be efficient to measure the intelligibility of speakers with speech disorders because it provides a way to metrically measure the difference between the distorted phonetic realization of linguistic units from the expected forms that occur in a normal speech.

We can also assume that such a process can be used in learning a foreign language (Kang et al., 2018), oral language acquisition or degradation due to aging.

6. Acknowledgements

This work was supported by Grant n°2014-135 from Institut National pour le Cancer (INCA) led by Pr Virginie Woisard at University Hospital of Toulouse and by Grant ANR-18-CE45-0008 from The French National Research Agency in 2018 RUGBI project ”Improving the measurement of intelligibility of pathological production disorders impaired speech” led by Jérôme Farinas at IRIT.

7. Bibliographical References

Astesano C., Balaguer M., Farinas J., Fredouille C., Gaillard P., Ghio A., Giusti L. et al. (2018), Carcinologic Speech Severity Index Project: A Database of Speech Disorders Productions to Assess Quality of Life Related to Speech After Cancer, LREC, 7-12 May 2018, Miyazaki (Japan)

Bechet F (2001), LIA_PHON : un système complet de phonétisation de textes, *Traitement Automatique des Langues* - *TAL* - Vol 42 n° 1 - pp 47-67, 2001

Clements G.N. (2005), The role of features in speech sound inventories In Raimy & Cairns, eds., Contemporary Views on Architecture and Representations in Phonological Theory. Cambridge, MA: MIT Press, p 19-68

Connolly J H. (1997) Quantifying target—realization differences. Part I: Segments, *Clinical Linguistics & Phonetics*, 11:4, 267-287, DOI: 10.3109/02699209708985195

Chomsky N., Halle M. (1968), The Sound Pattern of English. New York: Harper & Row

Dutrey C, Adda-Decker M, Yamaguchi N. Alignement de séquences phonétiques pour une analyse phonologique des erreurs de transcription automatique. *JEP-TALN-RECITAL 2016*, 2016, Paris, France. pp.46-54. ⟨halshs-01399054⟩

Ghio A, Rossi M (1995) Parallel distributed processes for speaker independent acoustic-phonetic decoding. International Congress of Phonetic Sciences (ICPhS), 1995, Stockholm, Sweden. pp.272-275. ⟨hal-01665248⟩

Ghio A. (1997). Achile : un dispositif de décodage acoustico-phonétique et d'identification lexicale indépendant du locuteur à partir de modules mixtes. PhD Thesis. Université d'Aix Marseille, 1997, ⟨tel-01663493⟩

Jakobson R., Fant G., Halle M. (1951), "Preliminaries to speech analysis", MIT Press, Cambridge.

Kang, O., & Ginther, A. (Eds.). (2018). Assessment in second language pronunciation. London ; New York: Routledge.

Kent R. (1992). Intelligibility in speech disorders. Amsterdam/ Philadelphia. John Benjamins.

Kondrak, G. (2003), Phonetic Alignment and Similarity, Computers and the Humanities 37: 273- 291

Meyer TK, Kuhn JC, Campbell BH, Marbella AM, Myers KB, Layde PM. (2004). Speech intelligibility and quality of life in head and neck cancer survivors. Laryngoscope. Nov;114(11):1977-81

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet : LEXIQUE™//A lexical database for contemporary french : LEXIQUE™. L'année psychologique, 101(3), 447–462. https://doi.org/10.3406/psy.2001.1341

Wagner RA, Fischer MJ (1974) The string-to-string correction problem, Journal of the ACM, 21(1) :168–173. DOI:10.1145/321796.321811