Balancing word lists in speech audiometry through large spoken language corpora

Annemiek Hammer[1], *Bart Vaerenberg*[2,5], *Wojtek Kowalczyk*[3], *Louis ten Bosch*[4], *Martine Coene*[1,2], *Paul Govaerts*[2]

[1]Applied Linguistics, VU University Amsterdam, The Netherlands
[2] The Eargroup, Deurne-Antwerpen, Belgium
[3]Leiden Institute for Advanced Computer Science, The Netherlands
[4]Department of Linguistics, Radboud University, Nijmegen, The Netherlands
[5]Laboratory of Biomedical Physics, University of Antwerp, Belgium

a.hammer@vu.nl, vaerenberg@otoconsult.com, wojtek@liacs.nl, L.tenBosch@let.ru.nl, m.m.r.coene@vu.nl, govaerts@otoconsult.com

Abstract

This paper describes a distance measure which estimates the distance between a language sample and a reference corpus with regard to graphemes, phonemes and the relation between them. The underlying assumption of this approach is that a language's phoneme distribution can be partially accessed via graphemes. The advantage of using such a measure in speech audiometry is twofold: (i) it may be applied to determine how representative existing word lists are with respect to the distribution of speech sounds in the target language of the test subject; (ii) it enables the audiologist to generate highly representative lists based on large corpora of languages for which broad phonetic transcription is lacking. In this paper the development of the *de novo* distance measure is described and demonstrated for Dutch. The technique itself however, is language-independent and has been applied successfully to 10 other EU-languages. As such, it paves the way to generating representative word lists as part of speech audiometric test batteries for any given language.

Index Terms: phonemic balance, speech audiometry, corpus analysis, distance measure.

1. Introduction

Perception of speech is a key aspect in effective oral communication. Hearing loss reduces speech perception abilities and jeopardizes participation in society. Therefore, one of the aims of rehabilitation device fitting (hearing aids or cochlear implants) is to improve speech perception. In audiological centres, the evaluation of speech perception skills, so-called speech audiometry, is standard practice.

In speech audiometry, speech perception is assessed by presenting word lists containing short CVC-words (e.g. *house*, *bug*) to the patient. An essential requirement in constructing word lists for speech audiometry is to select words that reflect the distribution of phonemes in the language – i.e. the speech stimuli have to be phonemically balanced [1]. A language's phoneme distribution reflects 1) the presence of all speech sounds in a language that are used to create differences in meaning, 2) the number of occurrences of each phoneme in the ambient language.

2. The problem

Establishing a phonemic distribution of a language requires a significant amount of linguistic information and statistical analyses. Ideally, phonemic distributions are drawn from large-scale spoken language corpora including broad phonetic transcription. However, one major limitation in establishing phoneme distributions is that most languages do not have such a phonetically transcribed corpus available. Regarding speech audiometry, the consequence of the limited availability of data is that for many languages, balanced word lists are completely lacking.

In the past working with orthographic input in phonetically based systems has been explored as a potential way to overcome the above sketched problem. For instance, orthographic transcripts have been used in training text-to-speech systems [2]. Results reveal that grapheme-to-phoneme mapping learning approaches yield better results than knowledge-based approaches including sophisticated linguistic knowledge (e.g. phonotactic, phonological and phonetic rules). Also, in the development of Automated Speech Recognition (ASR), the use of graphemes as the subword units has been explored [3,4] and context-dependent grapheme-based ASR systems have been shown to perform similarly as compared to phoneme-based ASR systems.

However, a drawback in this approach is that graphemes do not directly map to phonemes and that cross-linguistic differences exist regarding this mapping. However, even for orthographically deep languages (e.g. English) a correlation between graphemes and phonemes has been found [5]. This indicates that although grapheme-phoneme correspondences are not perfect, approaching a language's phoneme distribution via graphemes provides for many languages the only way to shed light on its phoneme distribution. In addition, this approach opens up for cross-linguistic perspectives in speech audiometry design.

The objective of this paper is to provide a measure that indicates the graphemic distance between a word set and the target language. The graphemic distance will be explored in more detail for Dutch but has been applied successfully to 10 other Indo-European languages.

3. Method and data

The approach taken to graphemic balancing is to present the graphemic distribution of the target language in a histogram and to compare this to a histogram presenting the grapheme distribution of a word list. The similarity between histograms indicates to what extent the word lists are representative of the target language in graphemic space. As shown in section 4, the degree of similarity can be quantified.

We explored spoken language corpora of 11 European languages (Dutch, for which the results are presented here, and 10 other EU-languages, including Danish, German, Greek, English, Spanish, French, Finnish, Italian, Portuguese and Swedish) to obtain language specific grapheme distributions. For each language, a so-called *fingerprint* was created. This fingerprint included calibrated frequency counts of word-initial graphemes and bigrams (sequences of two consecutive graphemes).

The analysis was based on the readily available EUROPARL corpus which comprises approximately 30 million words of spoken language. The corpus is based on the proceedings of the European Parliament from 1996 and is parallel, i.e. the transcripts are paired by translation. The EUROPARL corpus is available online from 2001 [6].

The EUROPARL corpus is limited to topics discussed in the context of the European Parliament, a particular constraint which might bias grapheme distributions. As such, our analysis included an additional large-scale corpus of Dutch to verify to what extent the phoneme distribution could be affected by the analyzed corpus. We analyzed the Corpus Spoken Dutch (Corpus Gesproken Nederlands, henceforth CGN). This corpus consists of contemporary Dutch as spoken in The Netherlands and Flanders by adult-speakers. The total size of the corpus is nearly 9 million words.

From all corpora, lexicons were built including all tokens of the language and their frequency of occurrence in the corpus. The number of types included in the lexicon for each language of the EUROPARL corpus is presented in Table 1. The number of tokens in the Dutch lexicon based on CGN is 103203.

Table 1. *Number of types in the lexicon for the 11 EUROPARL corpora of 11 European languages.*

Language	*No. of tokens in lexicon*
Danish	247,518
German	296,264
Greek	146,883
English	51,014
Spanish	111,676
French	78,076
Finnish	551,403
Italian	115,503
Dutch	176,153
Portuguese	95,197
Swedish	244,529

4. Graphemic distance measure

4.1. Establishing grapheme distributions

The lexicons were used to perform grapheme statistics. Rather arbitrarily, grapheme distributions can be either based on token counts - the number of overall occurrences of a grapheme in the corpus - or type counts - the number of unique occurrences of a grapheme in the corpus. Our statistical grapheme analysis involves a correction for type and token count. We established type (N_{type}) and token (N_{token}) counts for the initial phoneme and bigrams, calculated their geometrical means (called virtual counts (N_v) hereafter) as presented in Eq. 1 and converted them to frequencies (f) by dividing every virtual count by the sum of all virtual counts (of all graphemes from the lexicon) (see Eq. 2).

$$N_v = \sqrt{N_{token} * N_{type}} \qquad \text{Eq. (1)}$$

$$f = N_v / \sum_i N_v^i \qquad \text{Eq. (2)}$$

The grapheme statistics for each language included such frequencies of the word-initial graphemes and bigrams. These frequencies can be represented in a histogram, see Figure 1 for an example of Dutch word-initial graphemes.

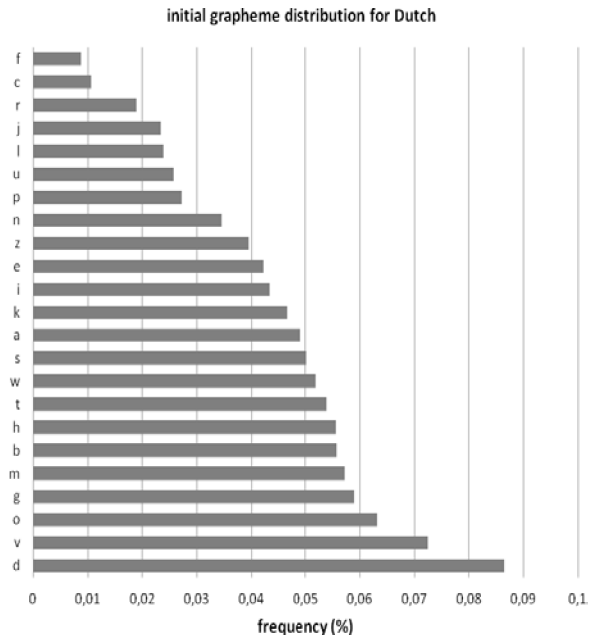

Figure 1. *Initial grapheme distribution based on Dutch CGN*

4.2. Establishing the z-distance

The grapheme distributions were calibrated through a random sampling procedure. Random samples (N 100,000) were drawn of 1000 'long' words, consisting of those words that are longer than the median word length in the language. As such, the actual length of the 'long' words that served as the basis for calibration purposes was dependent upon the target language.

For each sample the Euclidean distance (see Eq. 3) was used to calculate the distance between the normalized histograms for the initial phoneme and phoneme bigrams.

$$edist(x,y) = \sqrt{\sum_i (x_i - y_i)^2} \qquad \text{Eq. (3)}$$

Where *edist(x,y)* is the Euclidean distance between vectors *x=(x_1, ..., x_k)* and *y=(y_1, ..., y_k)*. In this way two vectors

of 100,000 distances were created, labeled dist_INIT and dist_BI.

These vectors were used to calculate the fingerprint parameters including the mean and standard deviation of dist_INIT and dist_BI. They were used to establish a distance measure that denotes the similarity between a list of words and the fingerprint of the target language in terms of standard-deviations from the mean. The resulting *z-distance* measure was defined as follows:

$$Z = \frac{\Delta_I}{2\sigma_I} + \frac{\Delta_B}{2\sigma_B} \qquad \text{Eq. (4)}$$

Where Δ_I is the Euclidean distance of a sample's initial character histogram to the fingerprint's initial character histogram, Δ_B is the Euclidean distance of a sample's bigram histogram to the fingerprint's bigram histogram, σ_I is the standard deviation of the Euclidean distance of the initial character histograms of 100,000 samples of 1,000 words taken from the fingerprint corpus to the fingerprint's initial character histogram and σ_B is the standard deviation of the Euclidean distance of the bigram histograms of 100,000 samples of 1,000 words taken from the fingerprint corpus to the fingerprint's bigram histogram.

5. Exploring the distance measure

5.1. Validating source of fingerprint

EUROPARL being composed exclusively of transcripts of topics discussed in the European Parliament, the obtained phoneme distributions could potentially be biased by the choice of the corpus itself and as such contaminate the fingerprints of the different languages. To control for this potential bias, we compared the outcomes of our analysis based on EUROPARL to those of another representative large-scaled corpora of Dutch (CGN) from which lexicons, grapheme distributions and fingerprints were drawn. The fingerprint parameters of the two representative corpora for the Dutch language are represented in Table 2.

Table 2. *Fingerprint parameters for Dutch in grapheme space.*

	EU Dutch	*CGN*
σ_I	0.0065	0.0067
σ_B	0.0011	0.0010

The results in Table 2 show that both fingerprint parameters can be compared to one another. This similarity is confirmed by a comparison between the normalized grapheme histograms (of initial grapheme and bigrams) obtained from the Dutch language dataset from EUROPARL corpus and from CGN. Specifically, we compared the CGN grapheme histogram (A) to the fingerprint generated from the EUROPARL corpus of Dutch (B) and vice versa. The results were as follows: from A to B we found a distance of 13.5 and from B to A of 14.3. These results indicate that the distance between fingerprints, drawn from different corpora, is small. As such, we conclude that the EUROPARL corpus is a valid source for the assessment of grapheme distributions.

5.2. Selecting a word sample: a demonstration

In view of the clinical purpose of the distance measure to generate representative word lists for speech audiometry, we would like to know whether these can be drawn from any text type or whether linguistic variation that comes with different rhetorical strategies (differences in lexicon, grammar and syntactic patterns, [7]) would yield differences in grapheme distribution as well. To answer this question, word samples were collected from different text types.

These included text types that are traditionally associated with five main rhetorical strategies: descriptions of an event in space (i.e. descriptions) or in time (i.e. narrations), clarifications of general concepts (i.e. expositions), instructions of future behavior to the reader (i.e. instructions) or argumentations [8]. For each category 300 samples of approximately 3,000 words each were collected.

All texts were segmented in words and only sets of unique types were retained. From this type set, the distributions of word-initial graphemes and bigrams were established and compared to the fingerprint of Dutch as in (6). This procedure yielded 1,500 z-distances that are presented in Figure 2. Significant differences were found between the grapheme distribution of text types and the fingerprint (X^2=37,155, p<.01). Post hoc analysis revealed that type sets built on instructional texts yield significantly higher distances to the fingerprints as compared to the other text types. This indicates that the graphemic distribution of types sets drawn from instructional texts is less representative for the target language as compared to type sets drawn from other text types.

Another interesting finding which can be read off from figure 2 is that variability is found between word sets obtained within one single text type. The variability is extremely large for instructional and narrative texts.

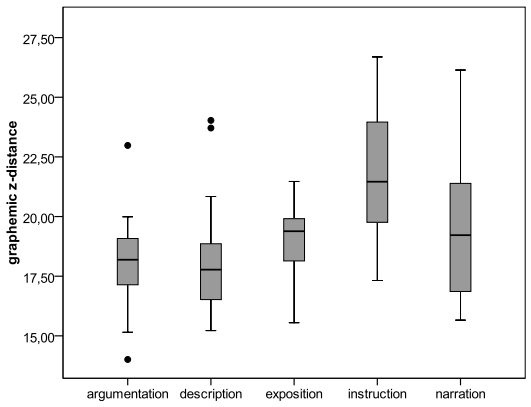

Figure 2. *Graphemic z-distances as a function of text type for Dutch.*

5.3. Comparison between word sets

Classical speech audiometry tests use pre-defined word samples for speech perception assessment. The word lists that are traditionally used in Dutch speech audiometry (NVA-lists) [9] were compared to the fingerprint of Dutch to obtain a graphemic z-distance. The words from the NVA-lists (12 words per list) are constructed from predefined phoneme sets

and, exceptionally, the lists are not phonemically balanced. These distances were compared to 10 word samples (of 20 words) that were randomly selected from the Dutch lexicon that yielded the best or worst z-distances. The results are presented in Figure 3. From this figure it can be observed that when the number of words is fixed at 20, variability of z-distances is expected to be between 47,0 (SD 0.35) and 82,3 (SD 1,4). The classic word samples have a mean of 96,8 (SD 2,6). Significant differences were found between sample class ($F(2,22)=2500,05$, $p<.00$). Post hoc analysis revealed significant differences between all categories.

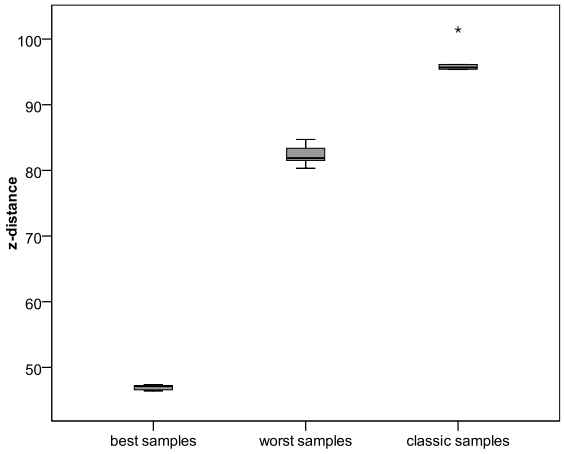

Figure 3. *Comparison between z-distances of best and worst word samples and the NVA-lists (classic word samples)*

6. Discussion and conclusion

In this paper we presented a measure that denotes the graphemic distance between any given sample of words and the target language. We have drawn grapheme distributions for 11 European languages based on the EUROPARL corpus. As shown for Dutch, this corpus can be considered representative to build graphemic fingerprints upon by comparison to fingerprints created from other large representative corpora (e.g. the Corpus Spoken Dutch). Finally, in this paper we also demonstrated the sensitivity of the distance measure in targeting the graphemic distance of a word sample to the fingerprint of the language.

The presented distance measure introduces two improvements in speech audiometry. First of all, the distance measure as presented in (4) operates language-independently, i.e. the measure can be used in every language which has a (written) lexicon available. Developments in the field of speech audiometry have been language specific so far and cross-linguistic perspectives in speech audiometry design are fairly new [10]. In this respect, the proposed distance measure opens up language boundaries in audiology and paves the way for comparisons in speech audiometry outcomes between speakers with different language backgrounds. Secondly, the measure could be of use to generate highly representative word lists for languages in which current speech audiometric test batteries use standard word samples. Some of them appear to be quite distant from and therefore rather unrepresentative for the target language's fingerprint (cf the results for the Dutch NVA lists in 4.3). They thus seem less appropriate for speech audiometry. In addition, such standard word lists are pre-recorded on disc and consists of very limited sets. In daily clinical practice patients are tested at a regular basis which potentially leads to memorizing the short CVC-words of which they are composed. Increasing the test sets by selecting representative word samples randomly is expected to reduce learning effects in speech audiometry and to increase test validity.

Finally, it should be observed that so far the analysis has been done on Indo-European languages alone. Additional research is needed to determine whether the observed findings can be generalized to typologically different languages, including tone languages such as Chinese or Japanese. Also, the distance measure denotes the distance between any word set and the target language in graphemic space. Future research will have to determine whether the observed distances in graphemes also carries over to the phonemic space. However, the distance measure can be applied to corpora including broad phonetic transcription, which would give an even better description of the phonemic balance of a word set. If these corpora become available in the future, the fingerprint and distance measure will become more sensitive in targeting phonemic balance of word samples. This seems of utmost importance to safeguard the method's possible success for languages with more complex orthographic systems, like English and French.

7. Acknowledgements

This research is supported by the European Community's Seventh Framework Programme FP7/2007-2013, funding number 262266 (OPTIFOX). The statistics of the eleven European languages are available on request by email.

8. References

[1] Hudgins, C.V., Hawkins, J.E., Karlin, J.E. and Stevens, S.S. The development of recorded auditory tests for measuring hearing loss for speech. Laryngoscope, 57, 57- 89, 1947.

[2] Van den Bosch, A., and Daelemans, W. Data-oriented methods for grapheme-to-phoneme conversion. EACL '93 Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics, 45 – 53, 1993.

[3] Doss, M.M., Stephenson, T.A., Bourlard, H., and Bengio, S. Phoneme-grapheme based speech recognition system. Workshop, 94 – 98, 2003.

[4] Doss, M.M., Rasipuram, R., Aradilla, G., and Bourlard, H. Grapheme-based automatic speech recognition using KL-HMM. Proceedings of Interspeech, 445 – 448, 2011.

[5] Coene, M., Hammer, A., Kowalczyck, W., Ten Bosch, L., Vaerenberg, B. and Govaerts, P. Quantifying cross-linguistic variation in phoneme-to-grapheme mapping. 14th Annual Conference of the International Speech Communication Association, August, 25-29 2013, Lyon, France.

[6] Koehn, P. Europarl: A Parallel Corpus for Statistical Machine Translation, MT Summit, 2005.

[7] Biber, D. Variations across speech and writing, Cambridge: Cambridge University Press, 1988.

[8] Werlich E. A Text Grammar of English, Quelle & Meyer, Heidelberg, 1976.

[9] Bosman, A.J., Wouters, J. and Damman, W. Realisatie van een cd vor spraakaudiometrie in Vlaanderen. Logopedie en Audiologie, 9:218 – 225, 1995.

[10] Wagener, K., Kühnel, V. and Kollmeier, B. Entwicklung und Evaluation eines Satztest für die deutsche Sprache I: Design des Oldenburger Satztests. Zeitschrift für Audiologie, 38: 1 – 32, 1999.