SYLLABIC STRUCTURE AND INFORMATIONAL CONTENT IN ENGLISH AND SPANISH

Vincent Aubanel, Chris Davis, Jeesun Kim

The MARCS Institute, University of Western Sydney, Australia
v.aubanel@uws.edu.au

ABSTRACT

This paper investigates the potential role of syllabic structure in characterising the informational content of running speech using an energy-based measure (the cochlea-scaled entropy, CSE index). We computed the CSE and compared how it aligned to the energy envelope for a corpus of English and Spanish sentences. We also compared these measures to syllabic structure, which differs markedly between the two languages. Results show that English exhibits a clear difference between informational and energy peaks in relation to the phonetic syllable nucleus, defined here in terms of the temporal mid-point of adjacent vowels. In contrast, in Spanish, both peaks align. Further, energy peaks occur later in the syllable in English, whereas they precede the nucleus in Spanish. Evaluation of internal syllable timing showed a more regular timing pattern in Spanish than English, which we suggest could have an implication for automatic selection of information bearing elements of speech.

Keywords: Speech Information, Speech perception, Syllabic structure, Cross-language comparison.

1. INTRODUCTION

Speech unfolds in time and meaning is built up by a complex aggregation of sequential speech sounds and silences. But not all speech sounds or silences make an equal contribution to meaning construction. For example, silences typically carry little phonetic information (not withstanding their potential pragmatic importance in a conversation), and word beginnings are more likely to inform the listener of what is being said compared to word endings (since, ambiguity decreases as more speech sounds becoming available, see for example [10]).

Focussing on the segmental level, several studies have proposed that vowels carry more information than consonants. The evidence for this has typically been obtained using a noise replacement paradigm in which differently categorized speech segments are replaced with noise and the effect on this on speech identification used as an index of information [4, 9]. Recently, a signal-based quantification of information in speech has been proposed, cochlea-scaled entropy (CSE) [14], one that has been claimed to outperform traditional vowel / consonant speech categorisation, while still bearing a close relation with vowel sonority. The development of this metric was inspired by information theoretic approaches [13] and on this view, information is defined in terms of local changes in the auditory transformed speech signal. That is, a region of speech in which acoustic energy levels change over time is argued to contain more information (and a higher CSE index) than a region where acoustic energy levels are more constant. Although it was developed using spoken American English, this metric can be applied to any language and it has been tested with Mandarin [8].

From the brief description above, one might be tempted to think that high-CSE regions will consist of highly changing speech regions such as those characterized by transitional phenomena (typical of stop consonants), however, owing to the auditory scaling that is implemented, changes in the lower part of the spectrum are given more weight over higher frequency regions, so high-CSE regions actually target vowels. Given this, it is interesting to explore the relation between information levels as measured by the CSE and the syllabic structure of speech. Recently, in an effort to pin-down the factors responsible for the advantage of selecting CSE-based regions in a noise replacement paradigm, a contrast between the CSE index and acoustic energy was conducted [2]. Here it was shown that there was a difference in alignment between CSE and the energy peaks, with CSE peaks occurring consistently earlier. Further, CSE peaks aligned with the syllable nucleus while energy peaks occurred later. The current study seeks to clarify the basis of this result by performing a cross language comparison with Spanish, a language which markedly differs from English in terms of syllabic structure. That is, in English, the morphological composition of words has a clear role in syllabic structure [3], whereas in Spanish (and other languages such as French) follows more closely the sonority hierarchy [6]. In particular, we

wanted to test whether the syllabic structure type had any influence on the alignment of the CSE and energy peaks with the syllable nucleus, as this could inform, through the analysis of the internal timing of the syllables, the relative informational flow in the two languages. Section 2 presents the two corpora under study, detailing their syllabic annotation and introduces the CSE metric as well as the energy measure used. Results are presented in Section 3, and implications for automatic characterisation of imformation in speech are discussed in Section 4.

2. METHODS

2.1. Corpora

English sentences were uttered by a female Australian speaker in her early twenties producing a 200 subset of the IEEE sentences [12]. Sentences were automatically aligned and manually checked and corrected. 180 sentences were retained for the current study, the same as the ones used in [2]. Phonological syllable boundaries were retrieved from the Cambridge Online Dictionary [1].

Spanish sentences were taken from a 180 subset sentences produced by the male talker of the Sharvard Corpus [1]. Forced alignment into phonemes was perfomed using EasyAlign [7] and 20% of the sentences were manually checked for correction. Phonological syllables boundaries were derived from orthographic syllable boundaries provided by the EsPal corpus [5] using a custom built Matlab script.

For both corpora, *phonological* syllable decomposition into onsets, nucleus and coda (O, N and C respectively) was determined automatically for each syllable by labelling vowels as N and any non-vowel as O if they occured to the left of the nucleus and C when they occured to the right. Additionally, aligned phoneme realisations for each sentence were grouped into successive vowel or consonant sound class clusters. This enabled to operationally define *phonetic* nuclei and edges of syllables as vowel and consonant cluster temporal mid-points respectively. Table 1 summarises the annotation done on the corpora.

2.2. Cochlea-Scaled Entropy and energy

The implementation of the CSE is defined as in [14] as the running sum of Euclidean distance d between successive 16 ms adjacent frames of auditorily-transformed speech spectra:

Table 1: Word and syllable counts in the English and Spanish corpora used in the study.

	English	Spanish
Sentences	180	180
Words	1410	1555
Phonological syllables	2424	1587
Phonetic syllables	4427	3188

$$(1)\quad d^2(t) = \sum_{f=1}^{F} [\rho(t+1,f) - \rho(t,f)]^2$$

$$(2)\quad CSE(t) = \sum_{k=-b/2}^{b/2} d(t+k)$$

where $\rho(t,f)$ is the output of an $F = 33$-channel roex filter [11] at time t and at frequency f, with centre frequencies covering the $26-7743$ Hz range, linearly-spaced on the equivalent rectangle bandwidth scale. b is the number of the adjacent frames over which to sum. Following [14] we use $b = 7$ (i.e., 112 ms), which corresponds roughly to mean vowel duration.

The metric for energy used here follows that used in [2], namely the root-mean-square (RMS) of the amplitude values over a frame size of 16 ms. The running sum of this quantity is taken as the final measure of energy, as for CSE (see Eq. 2).

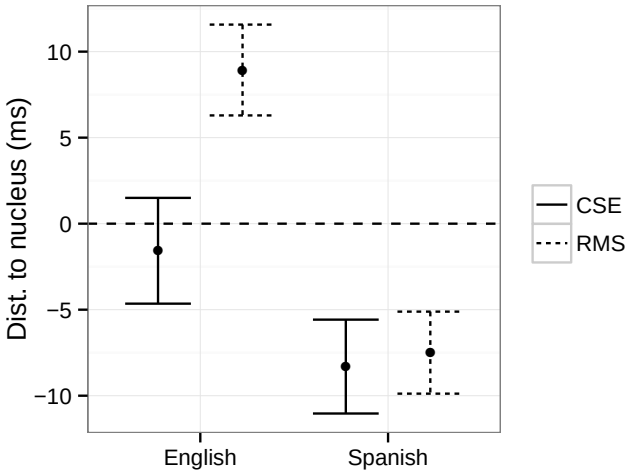
3. RESULTS

This section shows how CSE and RMS signals relate to the syllabic structure of English and Spanish, by examining the alignment of CSE and RMS peaks with phonetic syllable nuclei as defined in section 2.1.

In a first step, the first six 112 ms regions surrounding CSE and RMS peaks were incrementally identified (cf. [14]). Then, only pairs of CSE and RMS regions that overlapped were retained in order to compare their relative alignment to the syllabic structure. Finally, distance from the temporal midpoint of both CSE and RMS regions (i.e., the peak) and the closest phonetic syllable nucleus was calculated. Figure 1 shows the resulting distance for English and Spanish.

Individual t-tests on English and Spanish datasets show that for English, RMS and CSE peaks occur at significantly different points in time in relation to the phonetic syllable nucleus [Two sample t-test:

Figure 1: Distance to phonetic nucleus of CSE and RMS region peaks for English (N=1736) and Spanish (N=1500). Error bars show 95% confidence intervals.



$t(1695)=-5.08$, $p<.001$], with CSE peaks aligning with the syllable nucleus (mean$=-1.58$ ms) while RMS peak occuring 8.93 ms later on average. In contrast, for Spanish, both CSE and RMS peaks align at -7.9 ms before the syllable nucleus [One sample t-test: $t(1499)=-8.55$, $p<.001$], with the two peak time points being identical [Two-sample t-test, p=0.66].

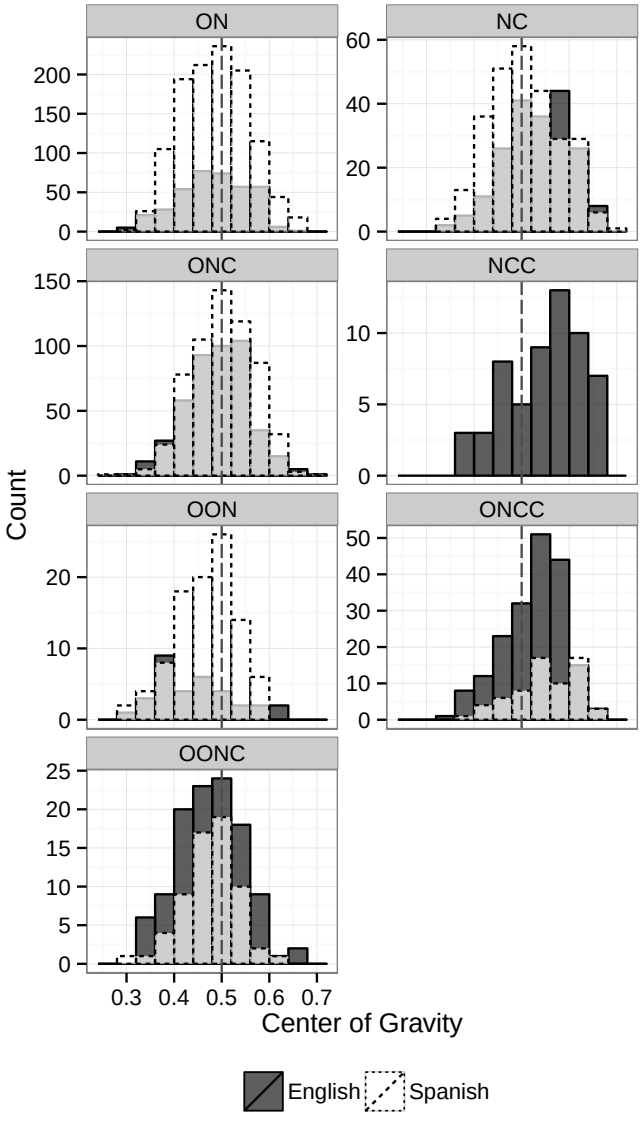
Two questions arise from this result. First, why is it that informational and energy peaks align in one language but not in another? The second question relates to the relative timing of the peaks: while CSE aligns with the phonetic syllable nucleus in English and RMS is positively shifted, both CSE and RMS peaks precede the nucleus in Spanish. As an approach to answer these questions, the phonological structure of syllables and their timing in the two corpora were examined. To this aim, the temporal center of gravity of each realisation of phonological syllables was calculated as:

$$(3) \quad CoG = \frac{1}{dN}\sum_{i=1}^{N} w_i t_i$$

where N is the number of segments in the syllable, w the weight associated with each segment, here all w_i are set to 1 (all segments are given equal weight) and t the temporal mid-point of each segment. d is the total duration of the syllable, so that the center of gravity is expressed as a value between 0 and 1. With this classical definition, a temporal center of gravity greater than 0.5 will represent a syllable with shorter segments towards its end. In contrast, a syllable having a sequence of short segments followed by longer ones will have a center of gravity lower than 0.5.

Figure 2 shows the distribution of the temporal center of gravity for the 7 most common syllable structures in English and Spanish, while Table 2 describes all syllable types encountered in the two corpora along with results of skewness tests.

Figure 2: Distribution of the temporal center of gravity of the different syllable types in English and Spanish. Note the difference in total counts across syllable types.



As can be seen on Figure 2 and Table 2, a number of notable differences emerge between the two corpora: first, while Spanish has a much greater number of simple syllable type (N, ON), English has more

Table 2: Syllable types and their occurrence in English and Spanish. Median of the temporal center of gravity is shown as well as Agostino test for skewness with its associated p value. Top panel: most common syllable types, also represented in Figure 2. Bottom panel: one-segment syllable (N) as well as less frequent syllable types.

	English				Spanish			
Syll. type	N	med.	skew	p	N	med.	skew	p
ON	380	.48	−.30	.016 *	1157	.48	.11	.133
NC	199	.54	−.32	.060	271	.50	.14	.326
ONC	450	.49	−.04	.708	599	.50	−.10	.308
NCC	58	.56	−.52	.087	0	-	-	-
OON	33	.44	.33	.383	98	.48	−.44	.065
ONCC	189	.53	−.49	.006 *	66	.55	−.50	.084
OONC	112	.48	.10	.640	64	.48	−.44	.127
N	99	.50	-	-	168	.50	-	-
OOON	2	.44	-	-	0	-	-	-
NCCC	1	.52	-	-	0	-	-	-
OONCC	30	.50	−1.04	.015 *	1	.46	-	-
ONCCC	13	.53	.06	.901	0	-	-	-
OOONC	14	.47	−.07	.890	0	-	-	-
OOONCC	3	.44	-	-	0	-	-	-
OONCCC	3	.43	-	-	0	-	-	-

composed syllable types (e.g., NCC). Second, English syllables ON, ONCC, and OONCC center of gravity show a significant negative skewness illustrating a 'late' center of gravity while no such skewness is observed in Spanish. This is confirmed by the median values of the temporal center of gravity, which on the main follows the phonological syllabic structure, with the center of gravity being displaced in the direction of the longest onset or coda.

4. DISCUSSION

Examination of the timing of syllables in both English and Spanish revealed a more symmetric pattern of the temporal center of gravity in Spanish as shown by an absence of skewness and a general alignement of the center of gravity with the temporal midpoint of the syllable. English data however displays more complex syllable types, and a more diverse patterning of center of gravity alignment, whose net effect is a global rightward shift of the temporal center of gravity. Taken together, this patterning of results provides a basis for the observed difference in terms of CSE and RMS alignement difference in English and Spanish.

It is interesting to note that while neither of the two CSE and RMS metric are tied to linguistic descriptions of the speech signal, their characterisation of the speech signal in terms of their local maxima shows a clear relation with linguistic constructs such as syllable composition. In particular, the alignment difference in English was found to be an explanatory factor in the performance difference observed by replacing CSE vs RMS regions by noise in a speech identification task [2], with CSE regions being slightly less disruptive than RMS regions.

On the basis of the current results one can hypothesise that given the alignment of CSE and RMS regions in Spanish, energy-based noise replacement would be just as effective in disrupting speech perception as one based on CSE. Further studies will attempt to validate this hypothesis, with potential important implications for signal-based characterisation of information in speech: while the CSE metric has been claimed to apply to languages universally, the current study suggests that the syllabic structure of the language could also play a role in determining information bearing portions of speech. This in turn has implications for designing algorithms to modify speech to enhance intelligibility, and in particular highlights the necessity to consider higher-level linguistic constructs such as word parts in understanding how people perceive speech, particular in challenging conditions.

5. REFERENCES

[1] Aubanel, V., García Lecumberri, M. L., Cooke, M. 2014. The Sharvard Corpus: A phonemically-balanced Spanish sentence resource for audiology . *Int. J. Audiology* 53, 633–638.

[2] Aubanel, V. and Faniad, H. and Kim, J. and Davis, C., in prep. Contributions of Cochlea-scaled entropy and energy in speech perception in noise.

[3] Borowsky, T. 1989. Structure preservation and the syllable coda in English. *NLLT* 7, 145–166.

[4] Cole, R. A., Yan, Y., Mak, B., Fanty, M., Bailey, T. 1996. The contribution of consonants versus vowels to word recognition in fluent speech. *ICASSP*. IEEE 853–856.

[5] Duchon, A., Perea, M., Sebastian-Galles, N., Marti, A., Carreiras, M. 2013. EsPal: One-stop shopping for Spanish word properties. *Behav Res* 45, 1246–1258.

[6] Fagyal, Z., Kibbee, D., Jenkins, F. 2006. *French: A Linguistic Introduction*. Cambridge University Press.

[7] Goldman, J.-P. 2011. EasyAlign: an automatic phonetic alignment tool under Praat. *Interspeech* Florence, Italy. 3233–3236.

[8] Jiang, Y., Stilp, C. E., Kluender, K. R. 2012. Cochlea-scaled entropy predicts intelligibility of Mandarin Chinese sentences. *164th Meeting of the Acoustical Society of America*. ASA 060006–060006.

[9] Kewley-Port, D., Burkle, T. Z., Lee, J. H. 2007. Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *J.*

Acoust. Soc. Am. 122(4), 2365–2375.

[10] Marslen-Wilson, W., Zwitserlood, P. 1989. Accessing spoken words: The importance of word onsets. *J. Exp. Psychol. Human.* 15(3), 576.

[11] Patterson, R. D., Nimmo-Smith, I., Weber, D., Milroy, R. 1982. The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold. *J. Acoust. Soc. Am.* 72, 1788.

[12] Rothauser, E. H., Chapman, W. D., Guttman, N., Hecker, M. H. L., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., Weistock, M., McGee, V. E., Pachl, U. P., Voiers, W. D. 1969. IEEE Recommended practice for speech quality measurements. *IEEE Trans. Audio Acoust.* 225–246.

[13] Shannon, C. E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27, 379–423, 623–656.

[14] Stilp, C., Kluender, K. 2010. Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. *P. Natl. Acad. Sci. USA* 107(27), 12387–12392.

[1] http://dictionary.cambridge.org/dictionary/british/. Last checked 31 Jan 2015