Translating Sound Adjectives by Collectively Teaching Abstract Representations

MARK CARTWRIGHT and BRYAN PARDO, Northwestern University

1. INTRODUCTION

Correctly translating adjectives that describe sound (e.g. “heavy”, “tinny”, “dry”, “soothing”) can be a difficult task [Zannoni 1997]. Resources such as the Oxford English Dictionary (OED) [Oxford University Press 1992], typically list the “audio sense” for only a small subset of the words commonly used to describe sound. For example, “warm” is a very commonly used sound adjective and the OED does not mention the audio sense. Directly translating the predominant (i.e. first) sense of a sound adjective into another language often results in an incorrect translation. For example, when “a warm sound” is typed into Google Translate [Google Inc. 2014], it responds with “un sonido cálido.” While the word-for-word translation is correct, the appropriate translation to correctly express the meaning is “un sonido profundo.” The word-for-word translation to English of “un sonido profundo” is “a deep sound,” not “a warm sound.” As a result, people relying on current translation technology may fail to communicate while believing they have. This, for example, would make it difficult for an English-speaking audiologist to correctly diagnose hearing problems for people whose primary language is not English.

Here, we describe a system that builds a translation map between sound adjectives of two languages: English and Spanish. This map is built from the collective intelligence of hundreds of participants who teach the system sound adjectives by indicating how well example sounds embody the adjectives. When two words are both strongly embodied by the same sound examples, they are considered synonyms. When the two words come from different languages, we consider one a translation of the other. The more frequently a pairing between two words occurs, the more certain the translation.

2. COLLECTIVELY TEACHING THE SOCIALEQ SYSTEM

SocialEQ.org is a web-based application that learns an audio equalization curve associated with a user-provided audio descriptor. Described in more detail in [Cartwright and Pardo 2013], this system was designed as a tool to build a knowledge base of audio equalization concepts to be used in an intelligent audio production system that responds to the descriptive language of the user.

To teach the system an audio equalization descriptor, participants are asked to “enter a descriptive term in the language in which you are most comfortable describing sound (e.g. ‘warm’ for English, ‘claro’ in Spanish, or ‘grave’ in Italian), pick a sound file which we will modify to achieve the descriptive term, then click on ‘Begin’.” Once a participant selects a descriptive term and a sound file, they are asked rate how well each of 40 modifications of the audio file embodies the adjective. Each modified version of the audio file is modified to alter the relative boost/cut to each of 40 frequency bands, spaced in a perceptually relevant manner (ERB) [Glasberg and Moore 1990]. From these rated examples, the system uses the method from [Sabin et al. 2011] to learn the relative boost/cut to apply each frequency band to modify a new sound to make it better embody the adjective. The result is a 40-band equalization curve learned from *one* participant which we call a *user-concept* (e.g. “Janet’s concept for ‘warm’ ”). By comparing hundreds of user-concepts one can find words that show broad agreement (the equalization curves are similar across multiple users teaching the system the same word) within and between languages.

To build such a collection of user-concepts, we recruited participants through Amazon's Mechanical Turk. We had 887 participants who participated in a total of 2322 training sessions (one session per learned word). We paid participants $1.00 (USD) per session, with the possibility of up to a $0.50 bonus, determined by the consistency of their responses. Out of 40 examples in each training session, 15 were repeats, to let us determine consistency of responses. Of the 2322 training sessions, 983 of them were contributed by participants recruited in English and used a version of the SocialEQ system in which all of the instructions were in English. The other 1339 sessions were contributed by participants recruited in Spanish who used a version of SocialEQ with instructions in Spanish.

We used the same inclusion criteria as specified in [Cartwright and Pardo 2013] to remove user-concepts from inconsistent participants (repeated sounds were labeled very differently) and those who showed no effort (e.g. completed labeling 40 sounds in under 1 minute). This left 676 participants who taught the system in 1602 sessions. Of these, 923 were English and 679 were Spanish, resulting in 388 unique English and 384 unique Spanish words. The median number of words contributed per participant was 1. Table I shows the top 10 adjectives in each language ranked by *agreement score between participants* [Cartwright and Pardo 2013]. This is a function of how often a word was contributed N and the inverse total variance Σ of the learned equalization curves: $agreement score = \frac{log(N)}{trace(\Sigma)_{descriptor}}$.

Table I. Top 10 English and Spanish equalization descriptors ranked by *agreement score*

Rank	English word	Sessions	Agreement Score
1	tinny	8	0.294
2	quiet	5	0.188
3	deep	6	0.164
4	light	6	0.151
5	warm	64	0.139
6	loud	26	0.137
7	heavy	15	0.124
8	dark	8	0.122
9	bright	19	0.112
10	energetic	5	0.107

Rank	Spanish word	Sessions	Agreement Score
1	pesado	10	0.142
2	agudo	5	0.111
3	suave	23	0.100
4	bajo	5	0.094
5	claro	33	0.092
6	fuerte	18	0.083
7	dulce	10	0.080
8	tranquilo	13	0.068
9	profundo	6	0.065
10	frío	13	0.063

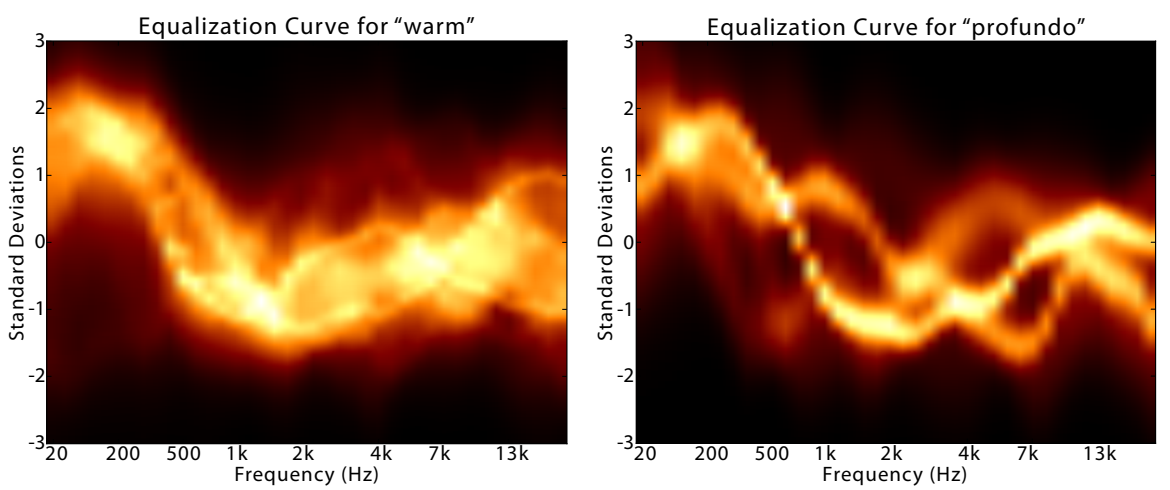

Fig. 1. Two closely related descriptor definition models: "warm" (N=64) and "profundo" (N=6), where N indicates how many people trained the system on the word in question.

3. REPRESENTING EQUALIZATION CONCEPTS

We combined individual's user-concepts into collective descriptor definitions by creating a mixture model. We model each user-concept as a Gaussian distribution, $\mathcal{N}(\mu_i, \Sigma_i)$, where μ_i is the learned equalization curve of the user-concept and Σ_i is a diagonal covariance matrix in which the variance for each frequency-band is set by $\sigma_{i,k}^2 = (\sigma_k - \sigma_k r_i)^2$ where σ_k is the sample standard deviation of frequency-band k for the equalization curves of *all* descriptors, and r_i is the ratings consistency for the session that learned user-concept i. Here we are using the ratings consistency as a measure of the uncertainty of the user-concept, mapping a consistency range of $[0, 1]$ to a per-frequency-band variance range of $[\sigma_k, 0]$. We then model each descriptor definition as follows:

$$P(x) = \frac{1}{N} \sum_{i=0}^{N-1} \mathcal{N}(\mu_i, \Sigma_i) \tag{1}$$

where $\mathcal{N}(\mu_i, \Sigma_i)$ is the distribution for the i^{th} user-concept.

Figure 1 shows the models of two descriptors. Here, the vertical dimension is the relative boost or cut of a frequency associated with that sound quality. The horizontal dimension is the frequency. Lighter values indicate a greater probability that a given boost or cut correlates with the descriptor.

4. MAPPING BETWEEN ENGLISH AND SPANISH

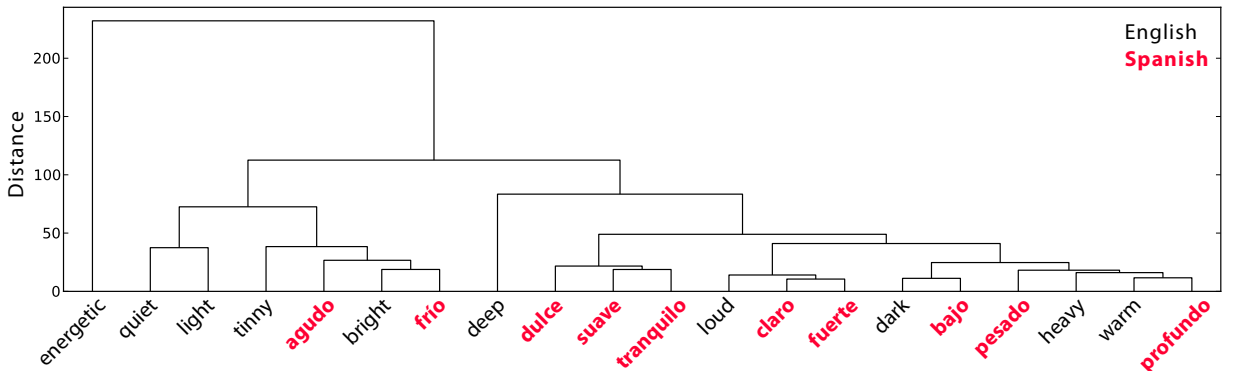

Fig. 2. Hierarchical clustering of the 10 descriptors with the highest agreement scores in each language.

Given two words represented, each represented by a mixture model (Figure 1), we can use a symmetric approximate KL-divergence [Hershey and Olsen 2007] as a distance measure between two words. To determine the relationships of the high-agreement words from Table I, we performed agglomerative hierarchical clustering using the "group average" algorithm [Hastie et al. 2001] and plotted the dendrogram in Figure 2. From this plot, we can see that the models displayed in Figure 1 ("warm" and "profundo") are closely related despite that "warm" typically translates to "cálido" in Spanish.

5. CONCLUSION

We presented a system to build an audio descriptor translation map between English and Spanish using data collected from hundreds of people. This provides an alternative to dictionary-based and statistical machine translation. This method of translating by collectively teaching intermediate abstract representations can potentially be extended to other domains, uncovering unknown relationships between languages. This work was supported by NSF Grant Nos. IIS-1116384 and DGE-0824162.

REFERENCES

M. Cartwright and B. Pardo. 2013. Social-EQ: Crowdsourcing an Equalizaiton Descriptor Map. In *Proc. of International Society for Music Information Retrieval, 2013*.

A. Disley and D. Howard. 2003. Timbral semantics and the pipe organ. In *Proc. of the Stockholm Music Acoustic Conference, 2003*.

A. Disley and D. Howard. 2004. Spectral correlates of timbral semantics relating to the pipe organ. *Speech, Music and Hearing* 46 (2004), 25–39.

C. Fritz, A. Blackwell, I. Cross, B. Moore, and J. Woodhouse. 2008. Investigating English violin timbre descriptors. In *Proc. of International Conference on Music Perception and Cognition, 2008*. 638–639.

B. Glasberg and B. Moore. 1990. Derivation of auditory filter shapes from notched-noise data. *Hearing Research* 47, 12 (1990), 103–138.

Google Inc. 2014. Google Translate. (2014). http://translate.google.com.

J. Grey. 1977. Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America* 61, 5 (1977), 1270–1277.

J. Grey and J. Gordon. 1978. Perceptual effects of spectral modifications on musical timbres. *The Journal of the Acoustical Society of America* 63, 5 (1978), 1493–1500.

T. Hastie, R. Tibshirani, and J. Friedman. 2001. *The elements of statistical learning*. Springer New York.

J. Hershey and P. Olsen. 2007. Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models. In *Proc. International Conference on Acoustics, Speech and Signal Processing, 2007*.

S. McAdams, S. Winsberg, S. Donnadieu, G. Soete, and J. Krimphoff. 1995. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research* 58, 3 (1995), 177–192.

O. Moravec and J. Stepanek. 2003. Collection of verbal descriptions of musical sound timbre in Czech language. In *Proc. of International Colloquium "ACOUSTICS'03"*. 23–26.

Oxford University Press. 1992. Oxford English dictionary. (1992).

A.T. Sabin, Z. Rafii, and B. Pardo. 2011. Weighting-Function-Based Rapid Mapping of Descriptors to Audio Processing Parameters. *Journal of the AES* 59, 6 (2011), 419–430.

M. Sarkar, B. Vercoe, and Y. Yang. 2007. Words that describe timbre: a study of auditory perception through language. In *Proc. of Language and Music as Cognitive Systems Conference, 2007*.

L. Solomon. 1958. Semantic Approach to the Perception of Complex Sounds. *The Journal of the Acoustical Society of America* 30, 5 (1958), 421–425.

L. Solomon. 1959a. Search for Physical Correlates to Psychological Dimensions of Sounds. *The Journal of the Acoustical Society of America* 31, 4 (1959), 492–497.

L. Solomon. 1959b. Semantic Reactions to Systematically Varied Sounds. *The Journal of the Acoustical Society of America* 31, 7 (1959), 986–990.

E. Toulson. 2003. A need for universal definitions of audio terminologies and improved knowledge transfer to the audio consumer. In *Proc. of The Art of Record Production Conference, 2003*.

A. Zacharakis, K. Pastiadis, G. Papadelis, and J. Reiss. 2011. An Investigation of Musical Timbre: Uncovering Salient Semantic Descriptions and Perceptual Dimensions. In *Proc. of the International Society for Music Information Retrieval, 2011*.

A. Zacharakis, K. Pastiadis, J. Reiss, and G. Papadelis. 2012. Analysis of Musical Timbre Semantics through Metric and Non-Metric Data Reduction Techniques. In *Proc. of 12th International Conference on Perception and Cognition*.

M. Zannoni. 1997. Approaches to translation problems of sensory descriptors. *Journal of Sensory Studies* 12, 3 (1997), 239–253.