INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

The role of linguistic and prosodic cues on the prediction of self-reported satisfaction in contact centre phone calls

Jordi Luque[1], *Carlos Segura*[1], *Ariadna Sánchez*[1], *Martí Umbert*[1], *Luis Angel Galindo*[2]

[1] Telefónica Research Edificio Telefonica-Diagonal 00, Barcelona, Spain
[2] Telefónica Móviles de España, S.A., Spain

jls@tid.es

Abstract

Call Centre data is typically collected by organizations and corporations in order to ensure the quality of service, supporting for example mining capabilities for monitoring customer satisfaction. In this work, we analyze the significance of various acoustic features extracted from customer-agents' spoken interaction in predicting self-reported satisfaction by the customer. We also investigate whether speech prosodic features can deliver complementary information to speech transcriptions provided by an ASR. We explore the possibility of using a deep neural architecture to perform early feature fusion on both prosodic and linguistic information. Convolutional Neural Networks are trained on a combination of word embedding and acoustic features for the binary classification task of ”low” and ”high” satisfaction prediction. We conducted our experiments analysing real call-centre interactions of a large corporation in a Spanish spoken country. Our experiments show that linguistic features can predict self-reported satisfaction more accurately than those based on prosodic and conversational descriptors. We also find that dialog turn-level conversational features generally outperforms frame-level signal descriptors. Finally, the fusion of linguistic and prosodic features reports the best performance in our experiments, suggesting the complementarity of the information conveyed by each set of behavioral representation.

Index Terms: Speech recognition, deep neural networks, customer satisfaction index (CSI), natural language processing

1. Introduction

Improving customer satisfaction is the cornerstone of strategy for contact centres and for any business development. Understanding customer needs, opinions, and expectations is becoming crucial for customer-centric enterprises. Usually, call centres (CC) serve as the primary customer-facing channel in many different industries. Agent Customer Service Representatives (CSR) interact with customers on behalf of an organization, becoming the key role and channel in maintaining brand reputation and customer experience and ensuring customer retention.

Substantially, customers now perceive a company through their interaction with CSRs in their centres. Dissatisfied customers are both more likely to share bad experience with others and prone to leave the company's service, yielding a need on understanding what drives satisfaction. Identifying and improving those key areas becomes critical and companies routinely have adopted numerous processes to enhance centre services and to detect organizational flaws. For example, it is very common to record calls for quality monitoring or acquire customer feedback after the call, aiming at increasing overall Customer Satisfaction Index (CSI).

Nowadays, current speech mining technologies open a window for information mining for business in call centre, by automatically performing the analysis on phone calls and transforming it into relevant information for both contact centres and their clients. Previous works have already performed automatic analysis and modeling of call-center conversations but using not speech related information or focusing on more traditional problems and tasks. In [1], we found a significant correlation between self-reported customer satisfaction and gender homophily, which may be considered for intelligent customer routing. The work in [2] performs automated topic detection for call taxonomy using transcripts provided by an Automatic Speech Recognition (ASR). In [3], a system is proposed for quality monitoring that combines speech recognition, pattern matching, and maximum entropy classification to rank calls according to the measured quality. Automatic understanding and retrieval of customer's intents is addressed in [4] using automatic transcriptions and Natural Language Processing (NLP) approaches. In [5, 6], authors took it as a emotion or sentiment classification task. The latter, proposing an opinion mining system aiming not only to detect if the customer is satisfied with the offered service but also to find out the intention of the caller. Finally, in [7], we proposed a feature learning approach from raw audio signal based on Deep Convolutional Neural Networks (DCNN) for continuous prediction of satisfaction and conflict detection.

In this work, we focus on the role of both language and discourse information in conjunction with acoustic correlates of emotion for the same customer satisfaction prediction task. We report our results on the individual contribution of both ASR transcripts and dialog turn-level descriptors together with prosodic cues on a corpus of real customer-agent calls interactions. Furthermore, we propose a Deep CNN based NLP approach that fuses linguistic and prosodic information at feature level. Results suggest that both sources of information are in fact complementary and that the CNN approach is able to successfully perform information retrieval of customer's sentiment.

2. Linguistic features

Linguistic features are obtained by an ASR system. It is trained in Spanish language using both manually annotated target Contact Centre data and external R&D data.

2.1. Automatic speech recognition

Target domain data is collected by CC partners within the EU BISON project [8]. Call contacts with user consent are recorded and both speech and call meta-data is anonymized by CC. Spanish call centre data were collected by Telefónica Móviles (TME) in 2016. The sampling of calls was performed taking into account call volumes from TME contact centres and the Spanish regions they serve in order to ensure a broad coverage of speaker and dialect variability. It accounts for a total of 22

http://dx.doi.org/10.21437/Interspeech.2017-424

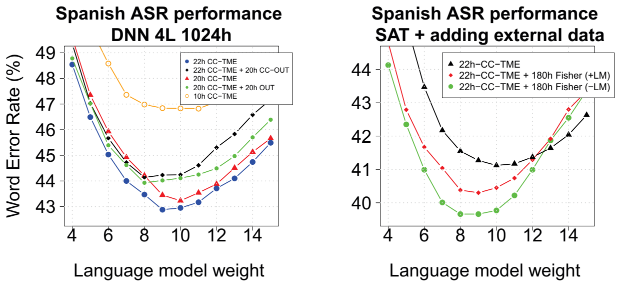

Figure 1: *Adding out-of-domain data to the target data (CC-TEF). (a) Fisher data for training only acoustic models (-LM) or for training both acoustic and language models (+LM). (b) "20h OUT" set corresponds to an internal telephonic non-conversational database [11] and "20h CC-OUT" corresponds to CC data from a different Spanish dialect.*

hours of speech. For assessment of the ASR performance, a total of 1.4 hours is kept for the test data set. In addition, licensed data from international organizations is employed for improving CC acoustic models and lexicon with a collection of different Spanish accents, dialects and pronunciations. SALA Chilean database [9], is used for lexicon augmentation. For improving acoustic modeling, the Spanish Fisher Speech Corpus [10] is employed.

The ASR is trained using the Kaldi toolkit [12]. It uses a single pass DNN system, with GMM pre-training, on top of filter-bank features. The speech is represented by 13-dim MFCC coefficients, plus their first and second derivatives and pitch features. The feed-forward DNN has 4 hidden layers (with 1024 neurons in each), not counting the output Softmax layer, and it is trained by mini-batch stochastic gradient descent. The GMM system makes use of discriminative feature transformations for GMM alignment. LDA and model-space adaptation using maximum likelihood linear transform (MLLT) are performed on top of triphone acoustic models [13], aiming to improve the separability of acoustic classes in the feature space. In addition, Feature fMLLR are also used for speaker adaptive training (SAT) of such acoustic models. The language model is trained using uniquely 22h of the transcribed speech from CC data, annotated and provided by TME. A business lexicon, also is provided by TME, which has been employed for augmenting the CC, Sala and Fisher lexicons together with products, brands, devices and more business-related words yielding to a dictionary around 50K words. Trigram language model is estimated using MIT Language Model Toolkit with Kneser-Ney Smoothing [14]. Figure 1 depicts the effect of adding the Fisher transcripts to the language model without interpolation weights. As a result, Fisher corpus is only used for acoustic training in this work. As linguistic knowledge is broadly available for Spanish as free-resource, the lexicon is obtained through the open source software SAGA [15], a phonetic transcriptor developed using linguistic rules for Spanish pronunciation at the Center for Language and Speech Technologies and Applications (TALP).

Speaker segmentation on the CC data is done by analysing independently each conversation channel, which is supposed to correspond to a single speaker. The results obtained by the DNN based system trained using both CC and R&D datasets, achieves 39.4% WER, see Fig. 1. The previous results suggest the benefits of incorporating more CC data compared to external data sources for both acoustic and language modeling. Also it is worth to mention, that techniques like speaker adaptation (SAT) and discriminative LDA-MLLR training have been incorporated to Bison TIDs systems, main difference between a) and b) 22h CC-TEF systems, black line in Fig. 1.

Contact centre data comprises spontaneous speech conversations with low recording quality, which makes automatic speech recognition (ASR) a highly difficult task. For typical call-centre data, even state-of-the-art large vocabulary continuous speech recognition systems produce a transcript with word error rate of 30% or higher [16]. Note that WER metric reported are directly computed using KALDI toolkit tools.

2.2. Natural language processing

We develop a Natural Language Processing (NLP) approach based on Bag-of-Words (BoW) modeling, broadly used in NLP and information retrieval tasks. Figure 2 depicts the main stages of such approach. Despite their simplicity, these models usually demonstrate good performance on text categorization and classification tasks so we decide to use this approach as a baseline system.

2.2.1. *Text normalization*

The phone call conversation is analyzed as a document hence no sentence splitting is needed. The system pools together the automatic transcriptions produced by the ASR from both conversation sides and performs a text normalization by removing low confidence words, pre-defined stop words and unusual terms. Therefore, those words recognized with a confidence lower than 0.4 by the ASR are discarded for further processing. Remaining words are tokenized and a corpus vocabulary is created. Note than no stemming nor lemmatization is performed.

2.2.2. *Vectorization and document term matrix normalization*

In order to represent documents in vector space, we create mappings from words to id vectors. We represent a set of documents as a sparse matrix, where each row corresponds to a call/document and each column corresponds to a word/term by using the vocabulary for constructing a document-term matrix (DTM) which accounts for word-co-occurence from call conversations, see Fig. 2. We transform DTM by normalizing term ocurrence as lengths of the conversations can significantly vary. This is done by Tf-idf, for word association to find a word's most relevant sub-category and create a reweighted DTM. We further reduce DTM dimensionality and also significantly improve accuracy by pruning the vocabulary and by projecting the DTM matrix into PCA components. We keep 10% of data variability by projecting the DTM into the first 100 PCA components. Note that the new DTM-PCA matrix has many fewer columns than the original DTM, usually yielding both accuracy

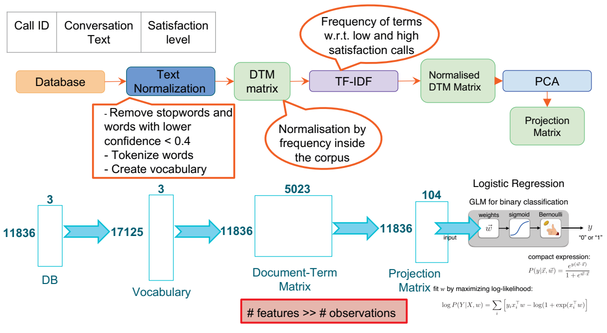
Call ID
Conversation Text
Satisfaction level
Frequency of terms w.r.t. low and high satisfaction calls
Database
Text Normalization
DTM matrix
TF-IDF
Normalised DTM Matrix
PCA
Projection Matrix
- Remove stopwords and words with lower confidence < 0.4
- Tokenize words
- Create vocabulary
Normalisation by frequency inside the corpus
3
11836
DB
17125
3
Vocabulary
11836
5023
Document-Term Matrix
11836
104
Projection Matrix
# features >> # observations
Logistic Regression
GLM for binary classification
weights
sigmoid
Bernoulli
$\vec{w}$
input
y
"0" or "1"
compact expression:
$P(y|\vec{x},\vec{w}) = \frac{e^{y(\vec{w}\cdot\vec{x})}}{1+e^{\vec{w}\cdot\vec{x}}}$
fit w by maximizing log-likelihood:
$\log P(Y|X,w) = \sum_i \left[y_i x_i^\top w - \log(1+\exp(x_i^\top w))\right]$

Figure 2: *Bag-of-Words based approach for modeling 'low' and 'high' self-reported satisfaction.*

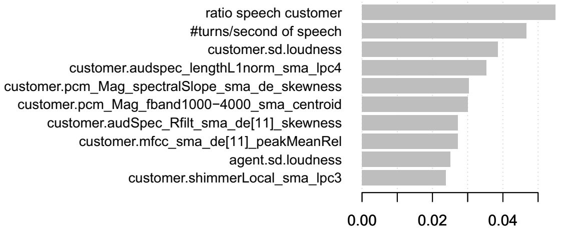

Figure 3: *Prosodic and conversational features ordered by importance given by Xgradient Boost modeling.*

improvement and lower training time.

2.2.3. DTM modeling

Once we have the Tf-idf matrix, or normalized DTM matrix, we use the glmnet package from R for fitting a generalized linear model (GLM), a logistic regression with elastic-net penalty. The regularization parameter λ that controls the overall strength of the penalty is estimated on training data using 10 fold cross-validation and a grid of values. The elastic-net penalty, α parameter, that controls the lasso (L1) and ridge (L2) penalty is also estimated by grid search.

3. Prosodic and conversational cues

The prosody and conversational cues relate to low-level descriptors (LLD) extracted from each audio recording on a frame basis. Speakers' turns are estimated based on a simple energy-based Speech Activity Detection on the corresponding conversation-side. From each speaker turn, acoustic LLD features are computed and post-processed. Such descriptors correlate with respect to speaker emotions, specially *F0* [17]. Handcrafted conversational cues which capture information at the customer-agent conversation level are also generated.

3.1. Acoustic cues and postprocessing

The fundamental frequency and speech loudness are extracted on a frame basis (using a 500ms window length and at 100ms rate). Articulation rate, defined as the number of syllable nuclei per phonation time, is also estimated using PRAAT [18], without requiring a transcription of the utterance. A sequence of articulation rate values is obtained by analyzing each turn.

Using the speaker turn segmentation, several values are computed for the previous acoustic features. First, the mean loudness and F0 value of each speaker turn is computed, using all loudness frames and *F0* frames between 62 and 600 Hz in order to avoid including error estimation values. From the sequence of mean values, its mean and standard deviation (sd) are computed. Besides, the sequence of *F0* mean values is normalized by the whole conversation mode so to compute the mean and sd of the $log2$ normalized values. Finally, *F0* frames are normalized by their mode and their $log2$ value is computed so to estimate their sd (total normalized sd). In addition, we used the OpenSmile toolkit [19] for extracting a bigger set of LLD descriptors using the paralinguistic 2013 configuration [20].

3.2. Conversational cues

While acoustic features focus only on a single speaker, conversational cues generally use data from both speakers in order to capture information at the dialogue level. Some examples of such features are the number of turns per second or the correlation value between the time series of acoustic and prosodic features obtained from each party side.

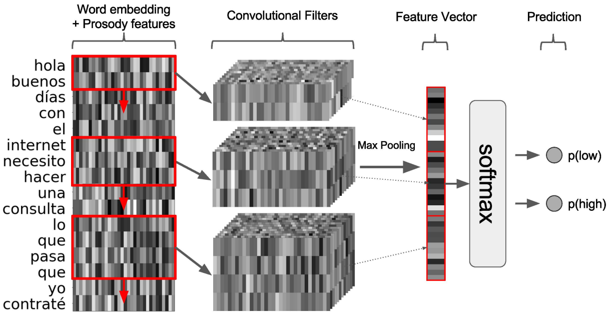

Figure 4: *CNN architecture using* 3 *sets of* 12 *filters of heights* $\{2, 3, 4\}$. *1-Max-pooled output from the conv. filters are concatenated into the feature vector and then fed to final softmax.*

3.3. Feature modeling

The prosodic and conversational cues together with the self-reported satisfaction scores have been used to train a gradient boosted regression tree model [21] using the scalable implementation in the *Xgboost* package [22] for R language. During training, the most relevant features are selected based on the information gain on the customer satisfaction index prediction. The 85 most relevant features are selected according to their contribution to the model. In Fig. 3 we present some of them and their relative contribution or gain to the model performance.

4. Neural Network approach

In this work we use a Convolutional Neural Network (CNN) over word embeddings [23] for classification in "low" and "high" satisfaction classes. We consider a one-layer CNN architecture similar to [24] due to its relative simplicity and good performance across many text classification tasks, making this model a modern standard baseline [25], [26],[27] .

4.1. Convolutional Neural Network Architecture

The automatic transcriptions are first filtered by setting low confidence words to out-of-vocabulary (OOV) symbol, tokenized and then converted to a *conversation matrix*, where each row stands for the word representation as shown in Fig.4. The word vector representations can be obtained from pretrained *word2vec* models [23], or they can be learned from scratch along with the network parameters. Since each row of the conversation matrix corresponds with a word, it is reasonable to perform the convolution only in this direction. This is accomplished by using 2D filters of height h and the same width as the word vector length. The height h controls the number of adjacent words that are jointly considered by the filter. The intuition behind CNN filters is that they capture features similar to n-grams, but they represent them more efficiently and in a

Table 1: *Example of sequence of words picked up by CNN. In the case of low satisfaction, the n-grams are related with problems in sound quality and issues with service, while for high class with politeness, agreement and payment extensions.*

Satisfaction	N-grams where ConvFilters peaked
Low	”nada no te escucho nada” ”estoy diciendo que pagu treinta” ”al al al” ”internet funciona sper mal” ”quiero poner un reclamo”
High	”gracias muy amable” ”necesito pedir una prrroga” ”igualmente muchas gracias” ”correcto ya perfecto” ”muchsimas gracias”

Table 2: *(Upper) Number of conversations per database for train/test sets and number of reported low/high satisfaction scores. (Bottom) AUC and F-score, between parenthesis.*

Database	Train (low / high)	Test (low / high)
Spain	494 (59 / 435)	210 (24 / 186)
Latam	11836 (1714 /10122)	5072 (734 / 4338)

	Dataset AUC, (F-score)	
	Spain	Latam
BoW-PCA	0.716 (0.324)	0.689 (0.262)
XGBoost (prosody)	0.58 (0.185)	0.610 (0.2398)
BoW-PCA + prosody (MW)	0.7309 (0.3420)	0.701 (0.269)
CNN	0.605 (0.212)	0.759 (0.410)
CNN + prosody	**0.733 (0.242)**	**0.772 (0.427)**

compact way. For instance, a single filter can match multiple n-grams at the same time, when the words composing those n-grams are very close in the embedding space.

The output feature map of each filter is downsampled using 1-max pooling, as in [28], which extracts just a single scalar from each feature map. In this case, the filters loose global locality information while keeping local information, a bit similar to the functioning of a bag of n-grams. Using this approach, one can find where each filter peaked in the conversation and do retrieval of n-gram that triggered the filter activation and its contribution to the final prediction, see Table 1.

Finally, 1-max pooled outputs from the filters are concatenated into the feature vector and then fed to final softmax layer for classification. Dropout [29] is applied to rows of the conversation matrix and to the feature vector with a keep probability of 0.85 and 0.5 respectively. The CNN was trained using categorical cross-entropy and Adam [30] optimization method, together with early stopping employing a 10% of the training data as evaluation. We performed a grid-search over the the network hyperparameters and selected a word embedding size of 32 together with 3 sets of 20 filters of heights $\{2, 3, 4\}$ for the vanilla network and heights $\{3, 4, 5\}$ for the fusion of linguistic and prosodic features. The CNN was trained using only Latam data and tested on both datasets described in Section 5.

4.2. Fusion of linguistic and prosodic cues

We propose a novel method for early fusion of linguistic and prosodic information, which consists of extending the embedding vector of each word with its corresponding prosodic features extracted from the audio. In this work we have run experiments concatenating the mean of F0 and loudness for each word. A weight and a bias parameter is also added that are jointly learned with the word vectors and the CNN parameters.

5. Experimental Evaluation

Table 2 reports both the distributions of train and test splits for each database and the distribution of CSI index. The Latin American data consists of a random sample of 16,908 inbound phone calls in Latin Spanish[1]. It was collected throughout one month such that comprises a huge variety of interactions between customer and representative. At the end of each call, the customer is called back and gently asked to complete a survey related to the service:

According to its previous call to our call centre, how satisfied, overall, are you with the telephone service of XXXX. Press 1-5 where 1 is very dissatisfied and 5 very satisfied.

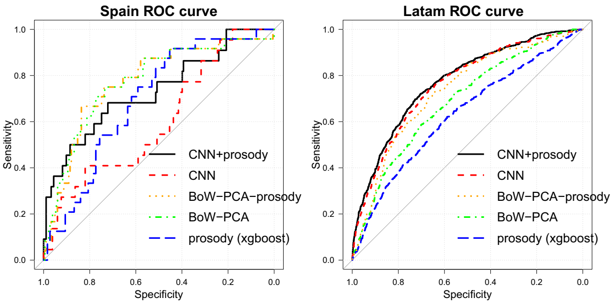

Figure 5: *ROC curves for different systems in both databases.*

We merged the customer satisfaction annotations into a binary decision: judgments 1 and 2 were taken as unsatisfied and 4 to 5 as satisfied. The classification task is designed based on those labels, namely to classify into low or high CSI. Similarly, the Spain database consists of 710 calls. In this case, the self-reported satisfaction ranges from 0 to 10. The low and high classes correspond to 0-1 and 9-10 scores, respectively.

5.1. Experimental Results

Table 2 depicts performances in terms of AUC and F-score for both Spain and Latam datasets. Figure 5 depicts the ROC curves. The scores from the linguistic features (BoW-PCA) are combined with the prosodic and conversational features at score level using Matcher Weighting (MW) fusion [31]. In MW fusion, scores are weighted by a factor inversely proportional to the Equal Error Rate, favouring the decisions taken by the most accurate classifier. Looking at the AUC values, the proposed approaches rank similarly for both databases. The combination of linguistic and prosodic features outperforms individual systems both in the MW fusion case and in the CNN approach. When predicting CSI without fusion, the linguistic features perform better. There is a mismatch in the results obtained by the CNN using only linguistic features in the different databases, probably due to the fact that the CNN was trained only on the Latam data. Noteworthy, the CNN+prosody approach is able to generalize even for a different dialect and variation of Spanish.

6. Conclusions

In this work we analyzed the role of linguistic, prosodic, and conversational information for customer's satisfaction prediction. For this purpose, we compared models trained on the individual contribution of ASR transcripts and dialog turn-level together with acoustic cues. Experimental results were reported on a corpus of real customer-agent calls interactions. Furthermore, we propose a Deep CNN based NLP approach that fuses linguistic and prosodic information at feature level. Results suggest that verbal communication convey more information than non-verbal with respect to customer's satisfaction and that both sources of information are complementary. Finally, we reported on a CNN-based system able to successfully perform information retrieval of customer's sentiment. We are currently extending the system for other CC related speech mining tasks and exploring in detail the fusion of acoustic features at word level.

7. Acknowledgements

This project has received funding from the EU's Horizon 2020 research and innovation program under grant agreement No 645323. This text reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains

8. References

[1] Q. Llimona, J. Luque, X. Anguera, Z. Hidalgo, S. Park, and N. Oliver, "Effect of gender and call duration on customer satisfaction in call center big data," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6–10, Dresden, Germany, Proceedings*, 2015.

[2] M. Tang, B. Pellom, and K. Hacioglu, "Call-type classification and unsupervised training for the call center domain," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, Nov 2003, pp. 204–208.

[3] G. Zweig, O. Siohan, G. Saon, B. Ramabhadran, D. Povey, L. Mangu, and B. Kingsbury, "Automated quality monitoring for call centers using speech and nlp technologies," in *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume: Demonstrations*, ser. NAACL-Demonstrations '06, 2006, pp. 292–295.

[4] M. Garnier-Rizet, G. Adda, F. Cailliau, J.-L. Gauvain, S. Guillemin-Lanne, L. Lamel, S. Vanni, C. Waast-Richard *et al.*, "Callsurf: Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content." in *LREC*, 2008.

[5] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Commun.*, vol. 49, no. 2, pp. 98–112, Feb. 2007.

[6] P. Li, Y. Yan, C. Wang, Z. Ren, P. Cong, H. Wang, and J. Feng, "Customer voice sensor: A comprehensive opinion mining system for call center conversation," in *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, July 2016, pp. 324–329.

[7] C. Segura, D. Balcells, M. Umbert, J. Arias, and J. Luque, "Automatic speech feature learning for continuous prediction of customer satisfaction in contact center phone calls," in *Advances in Speech and Language Technologies for Iberian Languages: Third International Conference, IberSPEECH 2016, Lisbon, Portugal, November 23-25, 2016, Proceedings 3*. Springer, 2016, pp. 341–345.

[8] BISON consortium, "Big speech data analytics for contact centers - bison," http://bison-project.eu, retrieved: 2017-03-1.

[9] A. Moreno, "SALA: SpeechDat Across Latin America," 2000.

[10] Graff, David, and et al., "Fisher spanish speech," *LDC2010S01. DVD. Philadelphia: Linguistic Data Consortium*, 2010.

[11] C. Torre, L. Hernández-Gómez, and D. Tapias, "CEUDEX: A Data Base oriented to Context-Dependent Units Training in Spanish for Continuous Speech Recognition," 1995.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi Speech Recognition Toolkit," 2011.

[13] R. A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2. IEEE, 1998, pp. 661–664.

[14] B.-J. P. Hsu and J. R. Glass, "Iterative language model estimation: efficient data structure & algorithms." in *in Proc. Interspeech*, 2008, pp. 841–844.

[15] A. Moreno and J. B. Mariño, "Spanish dialects: phonetic transcription." in *ICSLP*, 1998.

[16] J. Mamou, D. Carmel, and R. Hoory, "Spoken document retrieval from call-center conversations," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '06, 2006, pp. 51–58.

[17] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, March 2005.

[18] P. P. G. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, pp. 255–265, 2002.

[19] F. Eyben, M. Wollmer, and B. Schuller, "Openear - Introducing the Munich open-source emotion and affect recognition toolkit," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, 2009, pp. 1–6.

[20] B. S. et al., "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *in Proceedings of Interspeech, Lyon*, 2013.

[21] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[22] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 785–794.

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[24] Y. Kim, "Convolutional neural networks for sentence classification," *Proceedings of the Conference on Empirical Methods in Natural Language Processing, arXiv preprint arXiv:1408.5882*, 2014.

[25] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," *arXiv preprint arXiv:1510.03820*, 2015.

[26] A. Severyn and A. Moschitti, "Twitter sentiment analysis with deep convolutional neural networks," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015, pp. 959–962.

[27] R. Johnson and T. Zhang, "Semi-supervised convolutional neural networks for text categorization via region embedding," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 919–927.

[28] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 111–118.

[29] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[30] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference for Learning Representation (ICLR), San Diego, 2015*. arXiv preprint arXiv:1412.6980.

[31] M. Indovina, U. Uludag, R. Snelick, A. Mink, and A. Jain, "Multimodal biometric authentication methods: a COTS approach," *Proc. MMUA*, pp. 99–106, 2003.