Learning Voice Representation Using Knowledge Distillation For Automatic Voice Casting

Adrien Gresse, Mathias Quillot, Richard Dufour, Jean-François Bonastre

Learning Voice Representation Using Knowledge Distillation For Automatic Voice Casting

Adrien Gresse†, Mathias Quillot†, Richard Dufour, Jean-Francçois Bonastre

LIA, Avignon Université

lastname.firstname@†{alumni.}univ-avignon.fr

Abstract

The search for professional voice-actors for audiovisual productions is a sensitive task, performed by the artistic directors (ADs). The ADs have a strong appetite for new talents/voices but cannot perform large scale auditions. Automatic tools able to suggest the most suited voices are of a great interest for audiovisual industry.

In previous works, we showed the existence of acoustic information allowing to mimic the AD's choices. However, the only available information is the ADs' choices from the already dubbed multimedia productions. In this paper, we propose a representation-learning based strategy to build a character/role representation, called p-vector. In addition, the large variability between audiovisual productions makes difficult to have homogeneous training datasets. We overcome this difficulty by using knowledge distillation methods to take advantage of external datasets.

Experiments are conducted on video-game voice excerpts. Results show a significant improvement using the p-vector, compared to the speaker-based x-vectors representation.

1. Introduction

In order to broadcast to the widest audience, audiovisual companies target the market on an international, multilingual and multicultural level. At the same time, audiovisual creation producers pay more and more attention to the voices they attribute for a particular character or role in order to reinforce the audience's sense of immersion. Voice dubbing is one of the most important solution for audiovisual production localization and is capable of fostering this sense of immersion. Voice dubbing is about replacing the entire dialogs of the original creation by new voice-actors in the targeted language and cultural context. In this context, selecting the appropriate voices in a target language according to both the original voice and the role is a crucial task, referred as *voice casting*. Usually, a human expert called *artistic director* (AD) carries out the voice casting task in dubbing companies.

The major difficulty of voice dubbing lies in the fact that the "similarity" sought between an original voice and a dubbed voice is far from being a simple acoustical resemblance. It includes socio-cultural characteristics of both source and target languages and countries. Moreover, there is no well-established vocabulary for describing voices, characters and immersive effects. There are two limitations to the way ADs perform the voice casting task: the ADs are embedding their own socio-cultural characteristics in the casting task with the correlated subjectives biases, and 2), the ADs can't listen and memorize a very large set of voices. As a result, an AD usually works with a short list of actors he has listened to and/or with whom he has already worked.

Automatic tools able to measure the potential adequacy between an original voice in a source language/cultural context and a dubbed voice in a target language and context are of a great interest for audiovisual industry. They will help the ADs to remedy to the highlighted problems and to open the door for fresh voice talents, for example by preselecting a reasonable number of candidates inside a very large set of voices.

Perceptual voice similarity in the context of voice dubbing has been studied in [1, 2]. The authors show the importance of certain para-linguistic features (*e.g.* age, gender, speaker state, voice quality). In [3], the authors propose to estimate the "dubbing" proximity of two voices (one in the source language and one in a target language) using a i-vector/PLDA based speaker-recognition approach. Moreover, [4] makes the assumption that traces of the casting task performed by the ADs are present in the existing dubbed audiovisual creations. The proposed approach makes it possible to distinguish the *target* pairs of voices (*i.e.* a voice in the source language associated with the corresponding character voice in a target language) from *nontarget* ones (*i.e.* voices corresponding to different characters). A limit of this work is that the use of binary-supervised learning gives poor generalization capacities to the model, considering that interpolation could only be based on counter-examples.

Recent works in speaker recognition [5, 6, 7, 8] showed that deep neural network embeddings and end-to-end learning can outperform i-vectors. In this article, we propose to learn an original latent representation, called p-vector, from a character/role-based neural network approach. The p-vector representation should help the system to have better assimilation of the character dimension and consequently to better handle unknown voices. It constitutes the first contribution of this paper.

Nevertheless, a brake on the use of such a neural network approach is the need for a large amount of in-domain data, which is critical for many tasks, including the one we are dealing with in this work. The only information that we can use for a supervised learning approach is the operator's past voice selection from existing dubbing. In addition, voices used in our previous works come from a small number of characters. In this paper, we propose to remedy to this problem by applying knowledge distillation methods with the use of additional data, coming from a close domain, to extract the character/role specific information. More generally, we think the knowledge extracted, for example, from video-games could be transferred to other contexts, such as TV show voices characters.

This paper is organized as follows. We first present the approach and the generalized knowledge distillation framework in Section 2. Then we detail the corpus and we describe the experimental protocol we set-up in Section 3. We present our results and discuss them in Section 4. Finally, conclusions and perspectives are given in Section 5.

2. Approach

2.1. A character-based representation

In recent years, Deep Neural Networks have been proposed to learn task-oriented representational spaces allowing to disentangle the factors that explicate hidden data variability [9]. We propose to learn a dedicated representation called p-vector on professional acted voices. The p-vector space (p stands for "personnage" in French) is optimized on a character/role discrimination task. It learns to map the features space to a latent space that maximizes the character variability.

In general, input representation has a strong impact on the performance of machine learning applications. Here, we adopt the x-vector representation, originally introduced in automatic speaker recognition [8]. A large amount of data from many speakers are used to build the *speaker embeddings* space. Audio segments are projected into this space and characterized by x-vectors. x-vectors are seen here as a compact and fixed size representation of a variable length acoustic parameters vectors sequence. We make the hypothesis that the speaker embeddings contain entangled information corresponding to the character/role dimension. Hence, we propose to build a new representational space (p-vector) able to discriminate between the different characters.

2.2. Knowledge distillation

In the context of this work, we have to deal with relatively small number of data. We propose to use knowledge distillation in order to exploit additional data from a close domain to tackle this problem.

The generalized distillation framework [10] unifies two techniques that both introduce a teacher to guide a student model through its learning process. The first technique introduces the concept of *Privileged Information* [11] by adding a novel element $x_i^\star$ to the feature-label pair $(x_i,\, y_i)$ where $i \in [1...N]$, with N the number of samples. The second technique, referred as *Knowledge Distillation* [12], allows a simple neural network to solve a complicated task by distilling the knowledge from a cumbersome model. More generally, the teacher offers an opportunity for the student model to learn about decision boundary which is not contained in the training sample [10]. Typically, a *softmax* activation function computes the probability associated to every class $c \in [1...C]$ given the i-th sample with the following formula:

$$q_i = \frac{\exp\left(z_i/T\right)}{\sum_c \exp\left(z_c/T\right)}$$

where T refers to the temperature and z_i denotes the output computed for each class in the final layer. A higher value of T (> 1) gives a softer probability distribution over all classes. Distillation consists in raising the temperature until the teacher model produces proper soft-targets. The point is that soft-targets contain much more information than a simple one-hot encoding vector. These posterior probabilities, can be used as privileged information s_i to train a student model.

As illustrated in Figure 1, we fit the student model to the hard-targets (the one-hot encoded character labels) and soft-targets computed by the teacher model. This is achieved by minimizing the following loss:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left[(1-\lambda)l(y_i,\, q_i) \,+\, \lambda l(s_i,\, q_i)\right]$$

where l denotes the *cross-entropy* loss and s_i refers to the soft-targets from the teacher model. The λ parameter balances the imitation of soft-targets and hard-targets during the student model training.

The teacher-student framework has been used in different works [13, 14, 15, 16, 17, 18] for a wide variety of tasks such as noise-robust speech recognition, domain adaptation, and speaker normalization. The proposed approach originally extends this framework to acted voices and specifically to character/role representation.

Given the limited number of character labels in our corpus, we train the teacher model on an additional dataset with contains more character labels. We suppose it could help the student model to learn a robust, more general, representation by fitting to the soft-targets from the teacher. Also, we suppose the student model could bypass hard-targets limitation by simply ignoring them to some extent.

3. Experimental Protocol

3.1. Corpus

The voices from the *Mass Effect 3* role-playing game compose the main corpus. Originally released in English, this video-game has been translated and revoiced in other languages. In our experiments, we use the English and French versions of the audio sequences, representing 7.5 hours of speech in each language. Voice segments are 3 seconds long on average, each segment corresponds to a unique vocal interaction. Each English and French dataset contains $10,000$ voice segments. A character is then defined by a unique French-English couple of voice-actors. To avoid any bias in terms of speaker identity, we consider only a small subset of 31 different characters (13 female characters and 18 males), where we are certain that none of the actors plays more than one character.

To remedy the limited amount of characters in the *Mass Effect 3* corpus, we use additional data from another multilingual video-game called *Skyrim*. We limit this corpus to the English and French dialogues that are totalizing 120 hours of speech. We have $50,000$ segments in each language that are labelized according to 30 different character (7 females and 23 males). Since we do not have enough guarantee on the French-English correspondence of the segments and we are not sure that an actor plays a unique role, we do not use this corpus in the evaluation step. It only serves to transfer knowledge from the teacher to the student model in the distillation process. We make sure that there is no intersection between actors from *Skyrim* and *Mass-Effect 3* to prevent speaker-bias in the test set. Besides, all voice segments are high-quality studio-recorded audio files and we remove all segments shorter than one second.

3.2. Sequences extraction

We perform a usual acoustic parameterization of the audio signal that we transform into 60-dimensional feature sequences containing 20 MFCCs including the log of the energy plus the first- and second-order derivatives. We use a Hamming sliding window of 20ms with a 10ms overlap to compute the parameters. We perform a cepstral mean normalization and a voice activity detection to remove the low-energy frames that mainly correspond to silence. An x-vector extractor has been built with the Kaldi toolkit [19] and trained on the Voxceleb corpus [20].

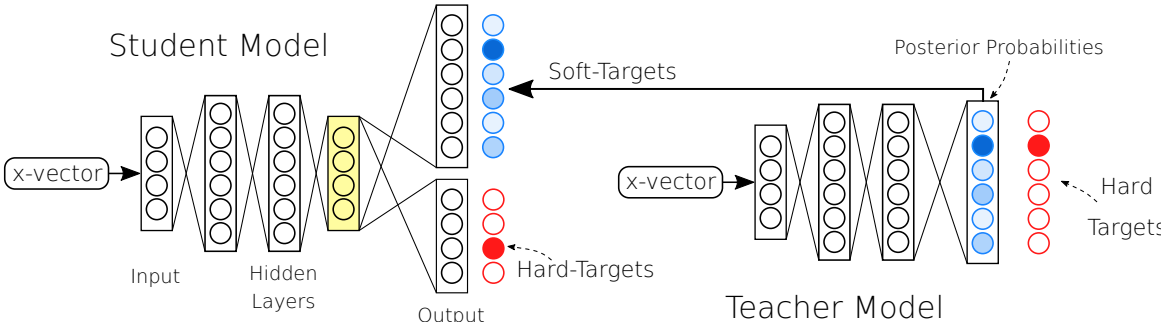

Figure 1: *The teacher model is trained to predict good soft-targets so that we can use them to train the student model. The teacher and student models can be trained either on the same or different corpus. The last layer of the student model (yellow) refers to p-vector.*

3.3. Training protocol

The quantity of voice segments in the *Mass Effect 3* corpus is not well balanced among the different characters because of their relative importance in the video-game. Consequently, we select only 16 characters that all have at least 90 voice segments from both English and French voice-actors. Segments are all randomly picked-out. Moreover, we create a k-fold cross-validation on this set of characters in order to have 4 of them in each fold. Thus, we have 4 distinct cases denoted A, B, C, and D that cover every character, each case involving 12 training characters and 4 characters kept-out for the evaluation. These 4 characters are completely unknown in the training part (they are not sharing any label or a speaker with one of the training data), making the voice-pairing task described in 3.4 extremely difficult. 20% of training data are used for validation. Regarding the additional corpus, we picked-out the same number of segments for all the 30 characters and we also divided it into two parts with the same ratio assigned to validation. As we said before, no data from *Skyrim* are used for the test.

Both teacher and student models are Multi-Layer Perceptron (MLP) created with Keras [21]. The two models begin with a similar network architecture. We connect a 512-dimensional input layer to two hidden layers of dimension 256. However, the teacher model ends with a single *softmax* function layer, while the student model has an extra hidden layer (*i.e.* corresponding to p-vectors) of dimension 64 connected to two different output layers (one for the soft-targets and one for the hard-targets). Hidden layers are all combined to a hyperbolic tangent activation. We apply a 0.25 dropout rate to the first two hidden layers and a 0.5 to the embedding layer. We use a *Xavier* initialization of the model parameters [22] and we use the *Adadelta* optimizer with its default configuration to solve the minimization of the *cross-entropy* loss function. Moreover, we use a batch size of 12 examples and we train the models during 300 epochs. We monitor the loss function on the validation set to avoid overfitting.

We first fit the teacher model to the features and labels from the additional dataset (*Skyrim*), considered as privileged information. The teacher model can be seen as a character/role discriminator. We compute the soft-targets given the *Mass Effect 3* features with the trained teacher. Then train the student model on input features, hard-targets and soft-targets (previously computed) from the *Mass Effect 3* corpus. The student model learns to fit the 12 hard-targets and the 30 soft-targets depending on λ which balances between soft- and hard-targets. It controls the weight of privileged information in the training process. Finally, p-vectors are extracted from the student embedding layer.

We perform the distillation process for varying temperatures $T \in [1...5]$. and we train the student modeling with different imitation values $\lambda \in [0, 1]$. Note that $T = 1$ denotes the absence of distillation and $\lambda = 0.0$ is equivalent to avoid the privileged information coming from the additional corpus. In opposition, when $\lambda = 1.0$ the model ignores the hard-targets.

3.4. Evaluation

Experimental validation is conducted on *Mass Effect 3* isolated characters. We evaluate the ability of the system to predict the choices made by the AD on the test voice segments.

To challenge the inherent quality of the learned representation, we first perform a clustering analysis using the k-means algorithm on the extracted p-vectors. We expressly set $k = 4$ to reflect the number of character labels present in the test set. Every voice segments then being gathered within the same cluster are assigned to the most represented character so that one cluster has one character label. Thus, a F-measure score is computed on the segment label hypothesis. Note that multiple clusters may be assigned to the same character, which constitutes a flaw. However, it remains a particular case indicating a bad result.

In addition, we evaluate the approach on a voice-pairing task using the similarity scoring system proposed in [4]. Here, we test the capacity to make a significant distinction between *target* (*i.e.* same character) and *nontarget* (*i.e.* different characters) pairs when we train the similarity model on p-vectors.

4. Results

Illustration 2 depicts the p-vectors space. We observe a good distinction between male and female characters. Same gender characters are also well separated, especially in case A and D. In C, we distinguish two clusters for the character "global_hackett" (*blue*) representing the two voice-actors playing this particular character. Their voices singularity might cause this distinction in the p-vectors space. We observe a similar phenomenon for character "hench_prothean" (*red*) in B.

4.1. Clustering analysis

Table 1 presents the results of the clustering analysis. For convenience we only present results corresponding to the extreme and central values of λ. The highest F-score (66% on average) involving the less variations (3%) among test cases A, B, C and D is observed with $T = 4$ and $\lambda = 1.0$. Unsurprisingly, baseline does not achieve good results, which is not surprising since x-vectors are designed to focus on the vocal identities of

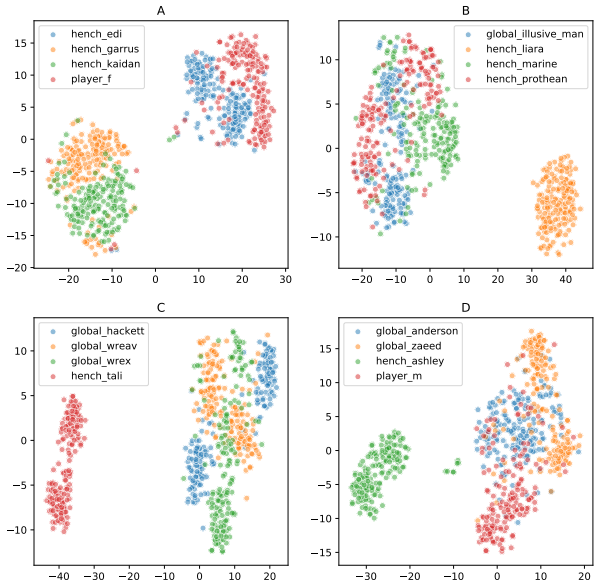

Figure 2: *Visualization of p-vectors from voice segments of case A, B, C and D using t-SNE. The colors discriminate the characters in each test case where English and French voices playing a same character share the same color.*

both English and French voice-actors more than on their character/role. Distillation has a positive impact given that higher temperatures have the better results, except for case B. However, according to the F-score (64%) obtained with the baseline on this case, we can argue that the acoustic proximity of male voices increases the confusions for the p-vector approach.

4.2. Similarity task

We evaluate the p-vector approach on the voice-pairing task presented in [4]. The system automatically attributes a score of similarity to every pair of voices (one voice in the source language, one in the target language). The similarity models is trained on the same corpus to avoid any bias. The objective of this task is to automatically distinguish *target* pairs (same character) from *nontarget* pairs (different characters). Table 2 presents the results of the Student's t-test performed between the scores of *target* and *nontarget* pairs. The t-score reflects the difference between the mean of the two groups. We observe a significant difference since all p-values are under the rejecting threshold. Moreover, the p-vector approach leads to an increased disparity between scores of *target* and *nontarget* pairs. The results presented in Tables 1 & 2 indicate that our approach offers better generalization on test voices than the speaker-oriented representation which shows that this new representation contains information dedicated to the character dimension.

5. Conclusion

In this paper, we introduced a deep neural network embedding called p-vector for automatic voice casting. The proposed approach firstly projects a speech recording in a character/role discriminant neural network representational space. It uses knowledge distillation methods to overpass data limitation problems. We use p-vector representation to apply character-based similarity metrics. We propose a very constrained protocol to counterbalance the limited amount of evaluation data. We observe a substantial improvement using our neural network embedding over the x-vector baseline. These results demonstrate that p-vectors contain information dedicated to the character/role dimension. We achieved to differentiate the same- and different-character pairs given the results of the similarity metric. Also, we successfully retrieved characters from unknown voices with a satisfying F-measure performance.

However, due to the limitations of our database and despite the rigorous protocol we designed, some caution should be taken. Confirmation of our findings on a bigger database with more character-labels is needed before to generalize to every kind of audiovisual production, character or language/culture.

The teacher-student framework allows us to compute new soft-labels and it could be more effective on larger training datasets with numerous character labels and multiple actors per label. Moreover, p-vectors allow the initiation of new research on the explicability/explainability questions, in particular in the context of artistic directors choices. We wish to confront p-vectors to a simple binary decision to observe the potential impact of a particular feature on the character dimension. Future work will replace the similarity system that discriminates between same- and different-character pairs with explanatory features (e.g. gender, voice-quality, timbre, prosody).

		A	B	C	D	Mean
x-vector baseline		50	64	54	47	53±06
p-vector						
	$\lambda = 0.0$	64	**76**	48	74	66±11
$T = 1$	$\lambda = 0.5$	55	63	48	62	57±05
	$\lambda = 1.0$	53	67	62	61	61±05
p-vector + distillation						
	$\lambda = 0.0$	52	57	54	75	60±09
$T = 2$	$\lambda = 0.5$	54	63	50	63	58±07
	$\lambda = 1.0$	54	66	51	72	61±09
	$\lambda = 0.0$	48	56	56	**76**	60±10
$T = 3$	$\lambda = 0.5$	63	64	51	71	62±07
	$\lambda = 1.0$	57	69	64	66	64±04
	$\lambda = 0.0$	53	63	58	75	63±08
$T = 4$	$\lambda = 0.5$	54	63	49	73	60±09
	$\lambda = 1.0$	63	66	63	71	**66±03**
	$\lambda = 0.0$	**70**	67	54	70	65±05
$T = 5$	$\lambda = 0.5$	52	55	54	73	59±08
	$\lambda = 1.0$	54	67	**66**	71	65±06

Table 1: *F-scores (%) computed on p-vectors with varying temperatures for test A, B, C and D. $T = 1$ indicate the absence of distillation. The λ parameter controls the soft-targets imitation.*

System	A	B	C	D	Cumulative
x-vector	5.60	42.50	4.69	33.77	26.55
p-vector	3.94	**44.62**	10.42	**82.11**	58.76
p-vector + distillation	**19.06**	30.28	**37.10**	32.27	**63.29**

Table 2: *Student's t-score obtained with the similarity scoring system. The last column considers the results from pairs of voices of all cases. Distillation parameters are $T = 4$ and $\lambda = 1.0$.*

6. Acknowledgment

This work is supported by the Digital Voice Design for the Creative Industry - TheVoice ANR project.

7. References

[1] N. Obin, A. Roebel, and G. Bachman, "On automatic voice casting for expressive speech: Speaker recognition vs. speech classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014.

[2] N. Obin and A. Roebel, "Similarity search of acted voices for automatic voice casting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 1642–1651, 2016.

[3] A. Gresse, M. Rouvier, R. Dufour, V. Labatut, and J.-F. Bonastre, "Acoustic pairing of original and dubbed voices in the context of video game localization," in *INTERSPEECH*, 2017.

[4] A. Gresse, M. Quillot, R. Dufour, V. Labatut, and J.-F. Bonastre, "Similarity metric based on siamese neural networks for voice casting," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[5] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014.

[6] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2016.

[7] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *INTERSPEECH*, 2017.

[8] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[9] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[10] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik, "Unifying distillation and privileged information," in *International Conference on Learning Representations*, 2016.

[11] V. Vapnik and R. Izmailov, "Learning using privileged information: similarity control and knowledge transfer." *Journal of machine learning research*, vol. 16, pp. 2023–2049, 2015.

[12] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015.

[13] R. Price, K.-i. Iso, and K. Shinoda, "Wise teachers train better dnn acoustic models," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2016, 2016.

[14] K. Markov and T. Matsui, "Robust speech recognition using generalized distillation framework." in *INTERSPEECH*, 2016.

[15] J. Li, M. L. Seltzer, X. Wang, R. Zhao, and Y. Gong, "Large-scale domain adaptation via teacher-student learning," 2017.

[16] S. Watanabe, T. Hori, J. Le Roux, and J. R. Hershey, "Student-teacher network learning with enhanced features," in *Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[17] T. Asami, R. Masumura, Y. Yamaguchi, H. Masataki, and Y. Aono, "Domain adaptation of dnn acoustic models using knowledge distillation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[18] N. M. Joy, S. R. Kothinti, S. Umesh, and B. Abraham, "Generalized distillation framework for speaker normalization," in *INTERSPEECH*, 2017.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[20] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[21] F. Chollet *et al.*, "Keras," 2015.

[22] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010.