MODELS OF TONE FOR TONAL AND NON-TONAL LANGUAGES

Florian Metze, Zaid A. W. Sheikh, Alex Waibel

Carnegie Mellon University
Language Technologies Institute/ InterACT
Pittsburgh, PA; U.S.A.

{fmetze,zsheikh,ahw}@cs.cmu.edu

Jonas Gehring, Kevin Kilgour, Quoc Bao Nguyen, Van Huy Nguyen

Karlsruhe Institute of Technology
Institute for Anthropomatics
Germany

{jonas.gehring,kevin.kilgour}@kit.edu
{quoc.nguyen,van.nguyen}@kit.edu

ABSTRACT

Conventional wisdom in automatic speech recognition asserts that pitch information is not helpful in building speech recognizers for non-tonal languages and contributes only modestly to performance in speech recognizers for tonal languages. To maintain consistency between different systems, pitch is therefore often ignored, trading the slight performance benefits for greater system uniformity/ simplicity. In this paper, we report results that challenge this conventional approach. We present new models of tone that deliver consistent performance improvements for tonal languages (Cantonese, Vietnamese) and even modest improvements for non-tonal languages. Using neural networks for feature integration and fusion, these models achieve significant gains throughout, and provide us with system uniformity and standardization across all languages, tonal and non-tonal.

Index Terms— Automatic Speech Recognition, Acoustic Modeling, Tone Modeling, Tonal Features, Neural Networks

1. INTRODUCTION

Tonal languages like Mandarin, Cantonese, and Vietnamese generally use tones to represent phone level distinctions [1], which are therefore essential to distinguish between words. Such tone information is generated by excursions in fundamental frequency, a feature that most recognition systems today discard as irrelevant for speech recognition. This state of affairs follows from a tradition of mostly processing non-tonal languages,[1] where pitch is assumed to be of value only at the sentential level, i.e. as one of three correlates of prosody (pitch, duration, intensity). Prosody, in turn, is considered only for supra-segmental modeling such as models of emotion, intonation, emphasis, discourse analysis, etc. [3, 4]. At the lexical level, pitch could contribute as a feature of lexical stress, but for a non-tonal languages such as English, only a small number of words are distinguishable by lexical stress alone (for example, INcline vs. inCLINE) [5]. Those few minimal pairs can often be distinguished by their time-frequency patterns as well. Indeed, the spectral patterns of speech have often performed sufficiently well even for tonal languages (e.g. Mandarin), so that pitch was and often still is ignored for all languages, in order to maintain system consistency and to avoid the extra complexities of pitch extraction (see for example the systems described in [6], which are using the same data that we are using here). Of course, dedicated efforts have obtained modest improvements using tone models and tonal features for recognition of tonal languages [7, 8, 9]. These however were usually tailored to the specifics of a particular language and its tone system.

In this paper, we wish to revisit these conventional notions, investigate more universal models of tone, and see to what extent they may help with the recognition of all languages. We will evaluate models of tone on several tonal and non-tonal languages. If they improve performance consistently for tonal-languages, and are not detrimental to non-tonal languages, tone could be included (rather than ignored) for all standardized speech recognition system builds, regardless of tonality, thus restoring the simplicity of a standardized approach. Our results show that this is indeed the case. Tonal modeling obtains considerable improvements for tonal languages, and modest improvements for non-tonal languages.

The paper is organized as follows: we begin by presenting the pre-processing we apply to extract pitch and pitch contours. We then explore where and how tonal features are to be merged in our system architecture: at the feature level, by detecting tones as tags, or by building tone dependent acoustic models. With the best of our models, we then construct systems with and without tonal models on two tonal languages, (Vietnamese, Cantonese) and two non-tonal languages (Tagalog and English), and discuss our results.

[1] There has been progress since [2], where the authors stated that "our strategy was to change the structure of our HUB4 English system only when absolutely necessary", but the spirit is certainly still present.

2. FEATURES FOR SPEECH RECOGNITION

In the past, a myriad of input features have been tested in ASR, and no single feature always outperformed all others. A "good" input feature was typically required to be simple to compute and stable to extract, exhibiting robustness to additive noise or other distortions, and having of course a good separability of distinctive properties of the speech signal. Human hearing of course operates under similar constraints; consequently, many feature extraction algorithms have, to some extent, been inspired by physiology.

In this paper, the baseline systems use either standard MFCC [10] features or Deep Bottleneck Features (DBNFs) as input features for Gaussian Mixture Models (GMMs). In extensive experiments on the BABEL [11] data-sets[2], this configuration was found to yield best results.

2.1. Minimum Variance Distortionless Response (MVDR) Spectrum

In addition to MFCC features, we also applied an MVDR [13] spectrum, to see how much combining multiple features helps on its own. In this work, we use twice-warped MVDR [14]. Stacking MFCCs and MVDRs at the input of a DNN was found to be helpful in bottleneck feature extraction for German Broadcast News [15] as well as exhaustive experiments as part of the NIST 2013 OpenKWS evaluation [16]. While MFCC and MVDR features are fundamentally similar and equally powerful, they are still complementary, and training a system on their union gives gains simply by increasing the robustness of the extraction.

Fundamentally different from spectral features, which capture the envelope of the speech signal, "pitch" features are typically used in addition to spectral features, and capture variations in the fundamental frequency of the speaker's voice.

2.2. Pitch Features

In so-called tonal languages, e.g. Cantonese and Vietnamese, (phonetic) pitch carries phonological (tone) information and needs to be modeled explicitly. To detect tones, one needs to detect rising, falling, or otherwise marked pitch contours. By themselves, pitch features are insufficient to distinguish all the phonemes of a language, but pitch (absolute height or contour) can be the most distinguishing feature between two sounds.

In this work, we extract pitch features using the approach described in [17]. We compute a Cepstrogram with a window length of 32 msec, and use dynamic programming to find the best path over time for the location of the maximum in these coefficients under certain constraints, like maximum pitch change per time unit. Additionally, we compute delta and double delta features using the three left and right neighbors as well as frame-based cross-correlation. This resulted in 8 additional coefficients (1 pitch, 6 delta and double-delta features and cross-correlation). These 8 coefficients were added to the original MFCC feature vector. A similar configuration was found to work best in [18].

[2]In this work, we used *babel101-v0.4c* (Cantonese), *babel106-v0.2f* (Tagalog), and *babel107b-v0.7* (Vietnamese). English experiments were run on an IWSLT task baseline [12].

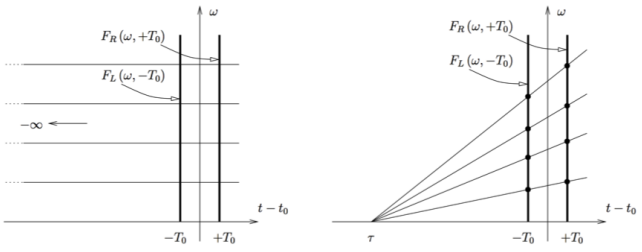

Fig. 2: Visualization of Fundamental Frequency Variation (FFV) features, from [19]: The standard dot-product between two vectors (spectra at a distance T_0 each from a center point) is shown as an orthonormal projection onto a point at infinity (left panel), while the proposed "vanishing-point product" for a point τ generalizes to the former when $\tau \to \infty$ (right panel).

2.3. Fundamental Frequency Variation (FFV) Features

In this paper, we report for the first time results of using FFV features [19] for speech recognition of tonal languages. FFV features have previously been used in tasks such as speaker verification. When compared to "standard" pitch-based features, their main advantage is that no explicit segmentation into speech and silence segments (for which pitch is not defined) is necessary.

Figure 2 shows a geometric interpretation of FFV features based on the "vanishing-point product": the change in F_0 between two short-term spectral vectors, which are at an equal distance from the analysis point, is modeled by computing the dot product for a range of values τ to the left, and to the right of the center. τ determines the amount by which the two spectra are warped before the dot product is being computed, so that the corresponding values lie on rays that intersect at the vanishing point, as in a painting using central perspective.

Rather than locating the peak in the resulting "FFV spectrum" (which is defined over $\tau \in [-\infty, \infty]$), we apply a filterbank, which attempts to capture meaningful prosodic variation, and contains a trapezoidal filter for perceptually "flat" pitch, two trapezoidal filters for "slowly changing" (rising and falling) pitch, and two trapezoidal filters for "rapidly changing" pitch. In addition, it contains two rectangular extremity filters, as unvoiced frames have flat rather than decaying tails. This filter-bank reduces the input space to 7 scalars per frame, which we use as additional "FFV" features in the final input vector.

(a) Late integration (tandem approach).

(b) Early integration as merged network input features.

Fig. 1: Late versus early integration of tonal features for acoustic models trained on deep bottleneck features. The non-filled network layers are only used during training.

3. MODELS OF TONAL INTEGRATION

Given the different roles that tone plays in different languages, a number of different approaches present themselves when integrating tones and tonal features into an ASR system.

3.1. Modeling of Tonal Phones

We examine two different modeling techniques for tonal phones. Our first approach involves modeling a phone's pronunciation and tone separately, which should allow for more flexibility when clustering context dependent quin-phones. We consider the tones to be tags that modify the base phones and therefore refer to this as the *tone tag* method. The Janus Recognition Toolkit (JRTk) [20, 21] allows these tags to be used as questions during the building of a decision tree. In Vietnamese, this results in a phone set containing 64 phones and 6 tone tags.

The second approach models each tonal variant of each tonal phone (vowels, diphthongs etc.) as an individual phone. This *tonal phone* approach increases the number of phones in Vietnamese to 205, but does not require any tone tags.

For example, in the tone tags model, the Vietnamese words $m\acute{a}$ *(mum)* and $m\tilde{a}$ *(horse)* have the same phone sequence, but different tags on the /aZ/ phone, /m (aZ tone_2)/ vs. /m (aZ tone_5)/, whereas the tonal phone model uses different phones, e.g. /m aZ_2/ vs. /m aZ_5/.

In Cantonese, we also examine the use of a demi-syllable based phone set with tone tags, where each Cantonese character (syllable) is decomposed into an initial-final (I-F) pair. These I and F units are called demi-syllables, with only F units requiring tone tags. We used 87 demi-syllables compared to only 46 phones.

3.2. Integration level

Our standard features (MFCC & MVDR) and the tonal features (FFV & Pitch) can be combined at different levels. In this section we compare a late integration setup using tandem features to an early integration setup using merged features for bottleneck feature training. We further investigate the effects of additionally using an ASR system with tonal features to train the context decision tree and write initial labels, which will shed light on how easy it is to bootstrap systems in new languages.

3.2.1. Late: As Tandem Features

Tandem features concatenate MLP-based features with standard ASR features once all feature extraction steps are finished [22]. Our tandem setup concatenates FFV and Pitch features with DBNFs [23] extracted from MFCC and MVDR features and stacks them over a 9-frame context. An LDA is used to reduce the feature dimensionality to 42. This setup allows us to add tonal features to an existing-DBNF based ASR system by simply retraining the acoustic model, keeping the number of parameters in the GMM constant.

3.2.2. Early: At the MLP Feature Level

Previous work has shown that training BNFs on a concatenation of multiple ASR features can result in improvements compared to BNFs trained on any of the individual features [15, 24]. As shown in Figure 1b, our merged DBNF training uses a 715-dimensional input vector consisting of 20 MFCC and 20 MVDR coefficients joined with 7 FFVs and 8 Pitch features and stacked over a 13-frame context window. The remaining DBNF topology and training procedure is unaltered.

3.2.3. Label Writing & Cluster Tree Construction using a Tonal System

So far, we have only modified our context dependent training to make use of tonal features. Both the labels and the context tree used in the context dependent training are created without any tonal features using a basic MFCC context independent system. In order to examine the effects of using tonal features at all stages of system training, we performed a second flat-start in Vietnamese where FFV and Pitch features are used alongside MFCCs from the start.

4. EXPERIMENTS

The Cantonese, Tagalog and Vietnamese systems are trained using data released within the IARPA Babel program [11], which consists of about 80 hours of transcribed conversational telephone speech per language. A 3-gram Kneser-Ney smoothed [25] language model is trained from the transcripts using the SRILM toolkit [26]. The acoustic models used in this work have been trained using GMMs with diagonal covariances and maximum likelihood; they are initialized using a flat start setup based on 6 iterations of EM-training and regeneration of training data alignments. Phonetic contexts are

clustered into roughly 10,000 context-dependent quin-phone states that also serve as targets for fine-tuning the DBNFs. The flat start and non-DBNF baseline systems use 13 MFCC coefficients that are stacked over 15 frames and reduced to 42-dimensional feature vectors with LDA.

The English baseline system is based on our 2012 IWSLT evaluation system [12], with the BNF-based acoustic model being replaced with by GMMs trained on DBNF features. In contrast to the other languages, this corpus contains wide-band Broadcast News speech.

The Cantonese, Tagalog and Vietnamese systems are tested on a 2-hour subset of their official 10-hour IARPA Babel development set and the English systems are tested on the 2012 IWSLT development set.

The "Baseline DBNF" systems use 20 MFCC coefficients concatenated with 20 MVDR coefficients, stacked over 13 frames [23, 15]. The deep bottleneck feature network consists of 5 layers containing 1,200 units each, followed by the bottleneck layer with 42 units, a further hidden layer and the final layer. Layers prior to the bottleneck are initialized with unsupervised pre-training as a stack of denoising auto-encoders [27]. Fine-tuning is performed for 15-20 epochs using the *newbob* learning rate schedule, which starts with a high learning rate that is exponentially decayed after the improvement in frame-level accuracy on a validation set falls below a fixed threshold. The activations of the 42 bottleneck units are stacked over a 9 frame context window and reduced to 42 features using LDA.

4.1. Experiments on Vietnamese and Cantonese

We trained both Vietnamese and Cantonese tonal systems in order to ascertain the best way to model tones and to identify how to optimally integrate the tonal features described above. As can be seen in Table 1, the tonal phone model consistently outperforms the tone tag model in Vietnamese by about 1%-1.5% absolute. The same is true for Cantonese (Table 2), where it is also significantly better than the demi-syllable system that models tones as tags.

By performing late integration, systems using tandem features improve on the baseline DBNF systems by 0.3%-0.7%. With improvements of 2.4% to 2.6%, integrating the tonal features early as additional input to the DBNF network performs much better. Using a tonal system as an initialization system for writing labels and building the cluster tree (denoted as Flat-Start with Tonal Features in Table 1) reduces the WER of the best systems by a further 0.4%.

In an initial set of experiments, we tried to determine the optimal combination of tonal and non-tonal features for the Vietnamese DBNF systems. We found that including Pitch features yields higher relative gains than including FFV features, but that their combination works particularly well (e.g. in combination with MFCCs: 55.7% WER with FFV, 55.1% with Pitch and 54.4% by including both). For the DBNF systems, we combined the tonal features with MFCC and MVDR features. Including MVDRs in the flat-start training resulted in worse performance, so we trained this system on MFCC, FFV and Pitch features.

System	Tone Tags	Tonal Phones
Baseline MFCC	70.3%	68.9%
Tonal features	66.8%	65.3%
Baseline DBNF	56.0%	54.7%
Tonal DBNF (late int.)	55.7%	54.0%
Tonal DBNF (early int.)	53.6%	52.1%
Flat-start with Tonal Features		
Tonal features	66.4%	65.1%
Tonal DBNF (early int.)	52.9%	51.7%

Table 1: Results obtained on Vietnamese in Word Error Rate (WER).

System	Tone Tags	Tonal Phones	Demi-syllable + Tone Tags
Baseline MFCC	66.6%	67.1%	66.6&
Tonal features	64.6%	63.8%	64.3%
Baseline DBNF	53.4%	51.4%	52.8%
Tonal DBNF (early)	52.5%	50.7%	52.5%

Table 2: Results obtained on Cantonese in Character Error Rate (CER).

4.2. Tonal Features in Non-Tonal Languages

We tested our best tonal setup on two non-tonal languages, Tagalog and English, in order to examine whether or not these potentially superfluous features had a detrimental effect. Table 3 shows that adding tone features actually results in small gains in these languages. In Tagalog, the DBNF system with early integration reduced the WER by 1.8% compared to the baseline DBNF system. Even for the non-Babel English system, a small improvement of 0.5% from 16.0% to 15.5% could be obtained with this approach. In contrast to the tonal languages, re-initializing the system with tonal features did not result in any gains. Again, the number of parameters in the GMM was the same for tonal and non-tonal systems.

5. DISCUSSION AND CONCLUSION

In this paper, we analyze the combination of multiple features for the recognition of multiple languages with different characteristics using deep neural network bottle-neck front-ends. We put a particular focus on tonal languages and features. We introduce Fundamental Frequency Variation features to ASR, and conclude that tonal features improve recognition in all conditions, when integrated using DBNFs.

Table 3 summarizes our results, and demonstrates a clear

WER/ CER (%)	ENG	TAG	CAN	VIE
Baseline	20.5%	69.0%	66.6%	68.9%
Best Baseline DBNF	16.0%	54.6%	52.8%	54.7%
Best Tonal DBNF	15.5%	52.8%	50.7%	51.7%
Δ (rel.) over Baseline	24.4%	23.5%	23.9%	25.0%
Δ over DBNF	3.1%	3.3%	4.0%	5.5%

Table 3: Summary of results obtained with bottleneck feature setups by merging tonal and non-tonal features.

benefit when fusing multiple, and tonal features when building ASR systems in tonal languages, without performing any language-specific modeling. Re-initialization of the acoustic model using a flat-start on tonal features is beneficial for tonal languages, but not for non-tonal languages.

In addition to observing gains by simply stacking features at the input layer of a neural network, we also observed gains by performing cross-adaptation of systems that had been trained with different tonal features in the case of Vietnamese: a multi-pass system in Vietnamese (developed for the OpenKWS 2013 evaluation [16]) benefited from using different tonal features in different adaptation steps (i.e. use Pitch in the first pass, but FFV in the second pass, etc.), with differences up to 4% relative between different adaptation methods. These effects indicate that further research into the choice of features is warranted.

Finally, the analysis of the number of questions for tone information in the context decision tree, as well as the dependency of final performance on the used initialization system, indicate that further improvements in WER (CER) can be expected by performing more systematic experiments on features and neural network based combination schemes.

6. ACKNOWLEDGMENTS

The authors would like to thank Kornel Laskowski for help in integrating the FFV features as well as Sarah Fünfer and Mirjam Simantzik for proofreading this paper.

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ ARL) contract number W911NF-12-C-0015. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ ARL, or the U.S. Government.

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575.

7. REFERENCES

[1] Z. Bao, *The Structure of Tone*. Oxford University Press, 1999.

[2] P. Zhan, S. Wegmann, and S. Lowe, "Dragon Systems' 1997 Mandarin Broadcast News System," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*. Lansdowne, VA; U.S.A.: NIST, Feb. 1998, http://www.itl.nist.gov/iad/mig/publications/proceedings/darpa98/pdf/h410.pdf.

[3] J. Pierrehumbert, "The phonetics and phonology of English intonation," Ph.D. dissertation, Massachusetts Institute of Technology, 1980.

[4] D. Hirst, "Lexical and non-lexical tone and prosodic typology," in *Proc. International Symposium on Tonal Aspects of Languages*, Beijing; China, Mar. 2004, pp. 81–88.

[5] A. Waibel, *Prosody and Speech Recognition*. Pitman Publishing, London, 1988.

[6] B. Kingsbury, J. Cui, X. Cui, M. J. F. Gales, K. Knill, J. Mamou, L. Mangu, D. Nolden, M. Picheny, B. Ramabhadran, R. Schlüter, A. Sethy, and P. C. Woodland, "A high-performance Cantonese keyword search system," in *Proc. ICASSP*. Vancouver, B.C.; Canada: IEEE, May 2013.

[7] S.-H. Chen and Y.-R. Wang, "Tone recognition of continuous Mandarin speech based on neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1995.

[8] C. J. Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny, and K. Shen, "New methods in continuous Mandarin speech recognition," in *Proc. EUROSPEECH*. Rhodes, Greece: ISCA, Sep. 1997.

[9] L. Lamel, J.-L. Gauvain, V. B. Le, I. Oparin, and S. Meng, "Improved models for Mandarin speech-to-text transcription," in *Proc. ICASSP*. Prague; Czech Republic: IEEE, May 2011.

[10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[11] Intelligence Advanced Research Projects Activity, "IARPA-BAA-11-02," http://www.iarpa.gov/solicitations_babel.html, 2011, last accessed July 7, 2013.

[12] C. Saam, C. Mohr, K. Kilgour, M. Heck, M. Sperber, K. Kubo, S. Stüker, S. Sakti, G. Neubig, T. Toda, S. Nakamura, and A. Waibel, "The 2012 KIT and KIT-NAIST English ASR systems for the IWSLT evaluation," in *Proceedings of the Ninth International Workshop on Spoken Language Translation (IWSLT)*, 2012.

[13] M. Murthi and B. Rao, "Minimum variance distortionless response (MVDR) modeling of voiced speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, 1997, pp. 1687–1690.

[14] M. Wölfel and J. McDonough, "Minimum variance distortionless response spectral estimation, review and refinements," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, 2005.

[15] K. Kilgour, T. Seytzer, Q. Nguyen, and A. Waibel, "Warped minimum variance distortionless response based bottle-neck features for LVCSR," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013.

[16] The National Institute of Standards and Technology, "NIST Open Keyword Search 2013 Evaluation (OpenKWS13)," http://www.nist.gov/itl/iad/mig/openkws13.cfm, Apr. 2013, last accessed: July 3, 2013.

[17] K. Schubert, "Grundfrequenzverfolgung und deren Anwendung in der Spracherkennung," Master's thesis, Universität Karlsruhe (TH), Germany, 1999, in German.

[18] N. T. Vu and T. Schultz, "Vietnamese large vocabulary continuous speech recognition," in *Proc. Automatic Speech Recognition and Understanding (ASRU)*. Merano, Italy: IEEE, Dec. 2009.

[19] K. Laskowski, M. Heldner, and J. Edlund, "The Fundamental Frequency Variation Spectrum," in *Proc. 21st Swedish Phonetics Conference (Fonetik 2008)*, Gothenburg, Sweden, Jun. 2008, pp. 29–32.

[20] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A One-pass Decoder based on Polymorphic Linguistic Context Assignment," in *Proc. Automatic Speech Recognition and Understanding (ASRU)*. Madonna di Campiglio, Italy: IEEE, Dec. 2001.

[21] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal, "The Karlsruhe Verbmobil Speech Recognition Engine," in *Proc. ICASSP 97*. München; Germany: IEEE, Apr. 1997.

[22] A. Faria, "An investigation of tandem MLP features for ASR," International Computer Science Institute, Tech. Rep., 2007.

[23] J. Gehring, Y. Miao, F. Metze, and A. Waibel, "Extracting Deep Bottleneck Features Using Stacked Auto-Encoders," in *ICASSP2013*, Vancouver, CA, 2013, pp. 3377–3381.

[24] C. Plahl, R. Schlüter, and H. Ney, "Improved acoustic feature combination for LVCSR by neural networks," in *Interspeech*, Florence, Italy, Aug. 2011, pp. 1237–1240.

[25] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–393, 1999.

[26] A. Stolcke, "SRILM – An extensible language modeling toolkit," in *Seventh International Conference on Spoken Language Processing*, 2002.

[27] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked Denoising Auto-Encoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.