A perceptual investigation of wavelet-based decomposition of *f0* for text-to-speech synthesis

Manuel Sam Ribeiro[1], *Junichi Yamagishi*[12], *Robert A. J. Clark*[1]

[1]Centre for Speech Technology Research, University of Edinburgh, UK
[2]National Institute of Informatics, Tokyo, Japan

m.f.s.ribeiro@sms.ed.ac.uk, jyamagis@inf.ed.ac.uk, rob.clark@ed.ac.uk

Abstract

The Continuous Wavelet Transform (CWT) has been recently proposed to model *f0* in the context of speech synthesis. It was shown that systems using signal decomposition with the CWT tend to outperform systems that model the signal directly. The *f0* signal is typically decomposed into various scales of differing frequency. In these experiments, we reconstruct *f0* with selected frequencies and ask native listeners to judge the naturalness of synthesized utterances with respect to natural speech. Results indicate that HMM-generated *f0* is comparable to the CWT low frequencies, suggesting it mostly generates utterances with neutral intonation. Middle frequencies achieve very high levels of naturalness, while very high frequencies are mostly noise.

Index Terms: speech synthesis, prosody, *f0* modeling, continuous wavelet transform, perceptual experiments

1. Introduction

Wavelets have been used in a variety of applications in Speech Processing for a number of years [1]. Recently, there has been a growing interest in the application of wavelets for the analysis and modeling of prosody in the context of statistical parametric speech synthesis. For example, wavelets have shown to be useful for the automatic annotation of prominence [2], as well as a pre-processing step for the modeling of *f0* in HMM-based synthesis [3][4], and voice conversion [5].

However, in most of these contributions [3][4][5], the Continuous Wavelet Transform (CWT) is used to decompose the *f0* signal into 10 scales, each approximately 1 octave apart. These scales are used to model *f0* without a clear understanding of their role in the overall signal. For example, [3] associate each pair of scales with a linguistically-motivated level (phones, syllables, words, phrases, utterance), although modeling still takes place at frame-level. In [4], the same decomposition is used, but the Discrete Cosine Transform (DCT) is used to model each scale at a supra-segmental level.

The general conclusion appears to be that approaches using signal decomposition tend to outperform approaches that do not. It was also seen in [4] that some scales are easier to predict than others, although their individual importance to the reconstructed signal is still not fully understood. Therefore, it is the goal of this work to explore the relevance of each of these wavelets scales (or frequencies) and their contribution of the perceived naturalness of speech.

We hypothesize that, on expressive datasets, short-term modeling approaches (such as MSD-HMM) tend to average most *f0* variation and generate a mostly neutral contour that is comparable to the low frequencies of the CWT. Approaches that use signal decomposition outperform these short-term approaches because, on top of the easily predictable low frequencies, there is always some improvements from the explicit modeling of middle-frequencies.

We assume a 10-scale wavelet-based decomposition of *f0*. Higher scales capture the CWT low frequencies and lower scales capture the CWT high frequencies. In this work, we index the decomposition by frequency, such that lower indexes correspond to low frequencies and higher indexes to high frequencies. Note that these these frequencies reflect the frequency of the wavelet component and are unrelated to the pitch range of the speaker. With this in mind, we propose to explore the following hypotheses in a series of evaluations:

- Listeners respond more to CWT middle frequencies (indexes 5 to 8) and associate them with higher levels of naturalness when compared to other CWT frequencies.
- Listeners do not respond much to the CWT low frequencies (indexes 1 to 4) and they achieve comparable naturalness to *f0* synthesized from an HMM-based system.
- High CWT frequencies (indexes 9 and 10) are mostly noise and do not contribute significantly to perceived naturalness.

To test these hypotheses, we run four perceptual experiments. In the first experiment we measure listeners' perception of selected wavelet ranges under a specific task. By asking participants to judge which word appears more prominent in an utterance, we test whether or not the wavelet scales are able to separate different prominence effects. This is a different approach to the task described in [2]. In the second experiment, we give listeners different utterances and ask them to judge whether or not they are similar in terms of naturalness. Considering the ratio of dissimilarity between wavelet scales, we use Multidimensional Scaling (MDS) [6][7] to establish a perceptual distance between all scales and natural speech. The final two experiments measures naturalness by asking listeners how much each utterance resembles natural speech. In the first of these, we run a traditional Mean Opinion Score (MOS) test, where participants are asked to rate an utterance on a scale of 1 to 5. In the second experiment, we run a MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test [8], in which listeners rate an utterance against a reference and against all other conditions. The key difference between the MOS and MUSHRA evaluations is that, in the first, participants rate an utterance without any reference. In the second test, participants are given a reference and are asked to judge each sample against it and against all conditions.

Section 2 introduces the CWT and section 3 details each condition and *f0* reconstruction. The following sections of the paper detail each experiment, and we conclude with a discussion of the results in section 8.

2. The continuous wavelet transform

A wavelet is a short waveform with finite duration averaging to zero. The continuous wavelet transform (CWT) can describe the *f0* signal in terms of various transformations of a Mother Wavelet. Scaling the Mother Wavelet, the transform is able to capture high frequencies if the wavelet is compressed, and low frequencies if it is stretched. The process is repeated by translating the Mother Wavelet.

The output of the CWT is an MxN matrix where M is the number of scales and N is the length of the signal. The CWT coefficient at scale a and position b is given by:

$$C(a, b; f; \psi) = a^{-1/2} \int_{-\infty}^{\infty} f(t)\psi(\frac{t-b}{a})dt \tag{1}$$

where f is the input signal and ψ is the Mother Wavelet.

3. *F0* reconstruction

The CWT is sensitive to discontinuities in the *f0* contour, so the signal was linearly interpolated over unvoiced regions. The interpolated log-f0 contour was then reduced to zero mean and unit variance, as this is required by the wavelet transform. To decompose f0, we use a decomposition approach identical to that described in [3] and [4], using 10 wavelet scales, each one octave apart. For reconstruction, we use a variation of the *ad hoc* reconstruction formula proposed by [3]:

$$f_0(x) = \sum_{i=1}^{10} w_i C_i(x)(i + 2.5)^{-5/2} \tag{2}$$

where scale 1 corresponds to the highest frequency scale and w_i is the weight given to scale i where $w_i \in \{0, 1\}$. Table 1 shows all experimental conditions with scales indexed by increasing frequency. These frequencies are related to the wavelet component and not the pitch range of the speaker. For each condition, *f0* is reconstructed from the wavelet domain zeroing out selected frequencies. For example, for condition *1-2*, the weight vector would be $\boldsymbol{w} = [0, 0, 0, 0, 0, 0, 0, 0, 1, 1]$.

Condition	Description	Freq. (Hz)
natural	Vocoded speech using natural parameters	-
all	All *f0* frequencies.	0.1-50
1-2	Low frequencies. Scales indexed at 1 and 2.	0.1-0.2
3-4	Low frequencies. Scales indexed at 3 and 4.	0.4-0.8
1-4	All low frequencies. Scales indexed at 1, 2, 3, and 4.	0.1-0.8
5-6	Middle frequencies. Scales indexed at 5 and 6.	1.6-3.2
7-8	Middle frequencies. Scales indexed at 7 and 8.	6.3-13
5-8	All middle frequencies. Scales indexed at 5, 6, 7, and 8.	1.6-13
9-10	High frequencies. Scales indexed at 9 and 10.	25-50
MSD-HMM	*f0* signal predicted from an MSD-HMM.	-

Table 1: Experimental conditions with approximate CWT frequency ranges.

4. Experiment 1: prominence

4.1. Data

For this experiment, we have recorded a native speaker reading the same utterance given different stimuli. All sentences consisted of 3 content words and had similar syntactic structure. The stimuli was chosen in order to suggest a different pitch accent location in the response. We have recorded 10 different utterances in 4 different contexts for a total of 40 utterances. Table 2 shows an example of stimuli and responses for one of the utterances.

stimulus	response
...	John won at Mary's.
Paul won at Mary's.	John won at Mary's.
John lost at Mary's.	John won at Mary's.
John won at Kate's,	John won at Mary's.

Table 2: Stimuli and responses for one utterance in the data set.

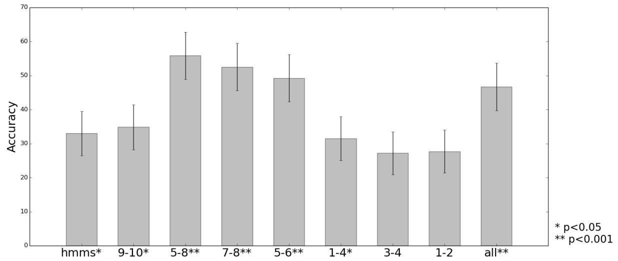

Figure 1: *Accuracy results.*

4.2. Design

For each of the 40 utterances, speech parameters were extracted using STRAIGHT [9]. When synthesizing the test data, all conditions except *natural* use mel-cepstral, aperiodicity, and duration parameters from the neutral response. This experiment relies on copy synthesis. It does not use synthesized parameters. The reconstructed *f0* contour was aligned with DTW at syllable level to the parameters from the neutral response. This ensures that listeners will not respond to durational or intensity cues when judging the utterances, as the only difference between them is *f0*. Utterances from the *natural* condition use all original parameters. A total of 400 unique utterances (10 sentences x 4 contexts x 10 conditions) were gathered. Each of the 25 participants listened to a randomized subset of 80 utterances. They were asked to select which of the 3 words appears more prominent or salient in each utterance, with the option to indicate that all words appear equally prominent.

4.3. Results

From the expected 2000 judgments (25 participants x 80 utterances), 23 were missing. This left us with an average of 198 judgments per condition, with each unique utterance having been judged either 4 or 5 times. To analyze the results, we used an approach similar to that described in [10]. That is, we considered the results from the *natural* condition as the gold set to which we measured the accuracy of all other conditions. Figure 1 shows the accuracy results from the experiment. Notes on the graph show the result of a binomial test assuming a chance accuracy of 25%. All conditions using the CWT middle frequencies achieve accuracy that is significantly above chance. Some conditions achieve a smaller, although significant effect, where we would expect not to see any. The reason for this might be how we are computing the results. When faced with uncertainty, listeners might default to an answer (that all words are equally prominent, for example).

5. Experiment 2: similarity

5.1. Data

To conduct the remaining experiments, we have used the freely available audiobook *A Tramp Abroad*, written by Mark Twain

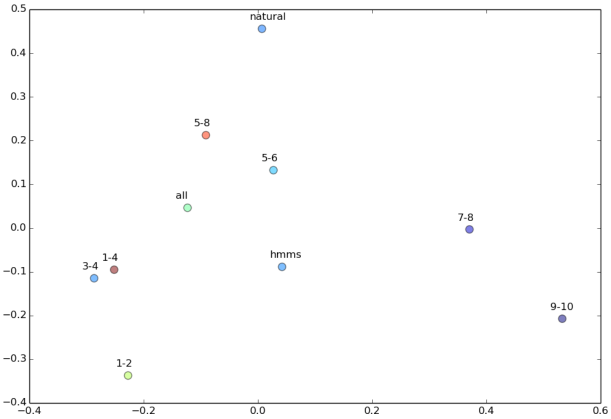

Figure 2: Two-dimensional representation of the dissimilarity matrix as estimated by MDS. Each point represents one condition and distances are representative of their dissimilarity in terms of naturalness.

and first published in 1880, available from *Librivox*[1]. Audiobooks are a rich source of speech data, as the reader mimics the voices of characters and attempts to convey some type of emotion depending on the circumstances. The data has been pre-processed according to the methods described in [11] and [12]. We have used a manually selected subset consisting only of narrated speech, thus setting aside direct speech data. The reason for this is that we intend to focus only on expressive read speech and avoid possible changes of speaking style and voice characteristics contained within the direct speech portions of the book.

A standard 5 state left-to-right HMM system was trained on roughly 5000 utterances. 20 utterances not in the training set were chosen for these experiments. Except for *natural*, all conditions used the same mel-cepstral, aperiodicity, and voicing parameters predicted from the HMM system. Duration is derived by force-aligning the data. The remaining conditions vary only *f0*, according to table 1.

5.2. Design

All 20 utterances were synthesized for each of the 10 conditions described in table 1. 10 native listeners participated in the experiment, each rating a total of 144 utterance pairs. Participants were instructed to listen to each pair carefully and judge if the pair is similar or different in terms of naturalness. Each pair given to the participants consisted of different utterances and different conditions. Within any three consecutive pairs, the same condition and utterance is not repeated. This prevents the task from being too easy and discourages participants from judging all comparisons as different[13][14].

5.3. Results

Considering the 45 distinct condition pairs, each pair was judged at least 32 times. A 10x10 dissimilarity matrix was constructed, indicating the fraction of times each pair was judged as different. Multidimensional Scaling (MDS) [6][7] was used to embed the dissimilarity matrix into a 2-dimensional space. The Euclidean distances between points in this space is representative of their perceptual distances. We have used the function *mdscale* from the Matlab statistic toolbox with Kruskal's normalized stress1, with a stress value of 0.086. Figure 2 shows the two-dimensional representation of the 10 conditions as judged by the participants. Distances between points are representative of their dissimilarity in terms of naturalness. Listeners naturally clustered CWT low frequency conditions (1-2, 3-4, 1-4), high frequency conditions (7-8 and 9-10), and middle frequency conditions (5-6, 5-8). The condition with *all* CWT frequencies appears to be closer to the conditions using the middle ones, with a greater distance from *natural* speech, which was surprising.

6. Experiment 3: MUSHRA

6.1. Design

In the MUSHRA test, participants are asked to rate all conditions of the same utterance in parallel from 0 (very poor) to 100 (very natural). Each condition has one slider and listeners are given the *natural* condition as reference. This utterance is also included in the unlabeled conditions and participants are instructed to judge at least one utterance as completely natural. This fixes the high end of the scale and all conditions are judged in relation to this.

We use the same data described in section 5.1. 10 native listeners rated all 20 sets of 10 stimuli, each stimulus originating from the conditions detailed in table 1. The order of the stimuli was randomized for each participant.

6.2. Results

From the expected 200 sets, 48 were discarded due to the hidden reference being judged as less than completely natural. These were excluded from the analysis. Figure 3 illustrates the distribution of the remaining 152 sets for all utterances and participants. Listeners ranked the CWT middle frequencies higher than the CWT low or high frequencies, with the HMM generated *f0* being comparable to the CWT low frequencies. Figure 4 shows the confusion matrix for Bonferroni-corrected pairwise Wilcoxon sign rank test.

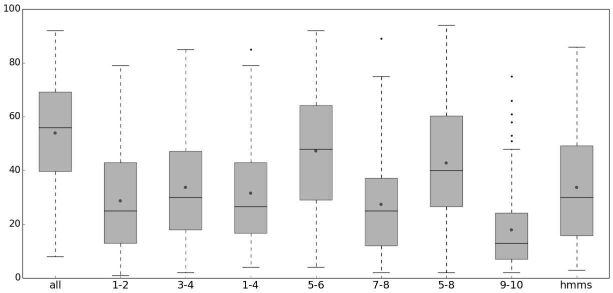

Figure 3: MUSHRA Test results.

	1-2	**3-4**	**1-4**	**5-6**	**7-8**	**5-8**	**9-10**	**hmms**
all	p<.001	p<.001	p<.001	p<.05	p<.001	p<.001	p<.001	p<.001
1-2		ns	ns	p<.001	ns	p<.001	p<.001	ns
3-4			ns	p<.001	p<.05	p<.01	p<.001	ns
1-4				p<.001	ns	p<.001	p<.001	ns
5-6					p<.001	ns	p<.001	p<.001
7-8						p<.001	p<.001	ns
5-8							p<.001	p<.01
9-10								p<.001

Figure 4: MUSHRA Tests - Bonferroni-corrected pairwise Wilcoxon sign rank test.

[1] http://librivox.org

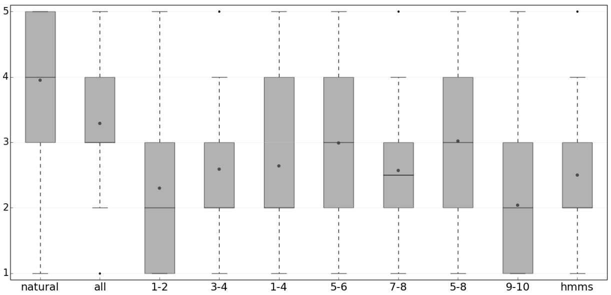

Figure 5: MOS Test results.

	all	1-2	3-4	1-4	5-6	7-8	5-8	9-10	hmms
natural	p<.001	p<.001	p<.001	p<.001	p<.001	p<.001	p<.001	p<.001	p<.001
all		p<.001	p<.001	p<.01	ns	p<.001	ns	p<.001	p<.001
1-2			ns	ns	p<.01	ns	p<.01	ns	ns
3-4				ns	ns	ns	ns	p<.01	ns
1-4					ns	ns	ns	p<.01	ns
5-6						ns	ns	p<.001	p<.05
7-8							ns	p<.05	ns
5-8								p<.001	p=.05
9-10									ns

Figure 6: MOS Tests - Bonferroni-corrected pairwise Wilcoxon sign rank test

7. Experiment 4: MOS

7.1. Design

In the MOS test, 25 participants were asked to rate each utterance on a scale of 1 (completely unnatural) to 5 (completely natural). No other instructions were given to the participants. Therefore, unlike the MUSHRA evaluation, participants have no reference against which to judge each utterance. All utterances were randomized for each participant. We use the same data described in section 5.1.[2]

7.2. Results

From the expected 1000 judgments, 1 was missing. Each unique utterance was judged 5 times and all conditions had 100 judgments, except 1 condition which had an utterance with 4 judgments and a total of 99 scores. Figure 5 shows a boxplot with the results for the MOS test. Listeners were quite conservative in the judgment of the natural speech, which has a median of 4. It still performs higher than the remaining conditions by the condition that reconstructs *f0* with all frequencies. Figure 6 shows the confusion matrix for Bonferroni-corrected pairwise Wilcoxon sign rank test. The ranking of the conditions is consistent with the one shown by the MUSHRA evaluation, which suggests that the two tests are quite similar. However, in the MUSHRA, participants are able to make direct pairwise comparisons using a larger scale, so differences between each condition are clearer.

8. Discussion and conclusions

The results from these evaluations support the initial hypotheses. We see evidence that native listeners tend to prefer the middle frequencies of the CWT when used to decompose the *f0* signal. The 5th and 6th scales consistently show higher degrees of naturalness when compared to the remaining scales. These have been previously associated with the word-level [3] [4].

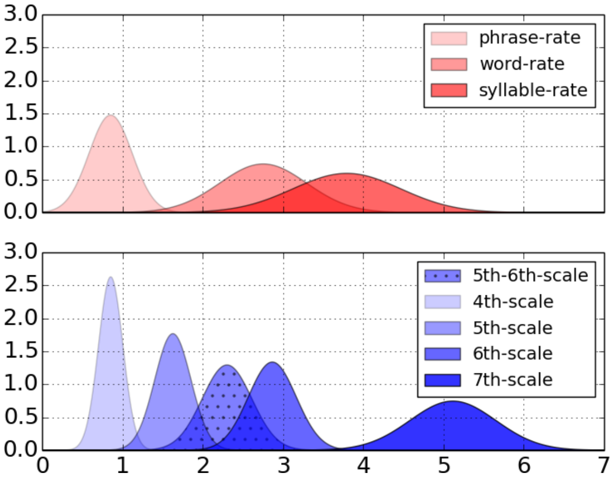

Figure 7: Unit (top) and peak (bottom) rate per second for selected units and scales.

Figure 7 shows distributions of word rates per second and peak (local maxima) rates per second in the CWT middle frequencies. Rates are computed at utterance level on 5000 utterances. Although these might change depending on speaker or speaking rate, the 6th scale matches the distribution of words, suggesting it could be modeled at word level. The 5th scale is best modeled at a level that is higher than the word. The distribution of intonational phrases could be associated with the 4th-scale. Therefore the 5th lies at a level higher than the word, but lower than the intonational phrase. This suggests that these middle frequencies contain most of the information that listeners associate with naturalness in expressive speech and might provide a suitable candidate when exploring supra-segmental models of *f0*. The 7th and 8th scale were also expected to rank higher than others, but these results show that they mostly resemble the higher frequencies. These have been associated with the syllable-level [3] [4], but figure 7 shows that the 7th scale does not match syllable rates.

Regarding the low frequencies, we observe that they behave as expected. In the MOS and MUSHRA evaluation, no significant differences were found between the ratings of all the lower scales. Similarly, we have failed to observe significant differences between these frequencies when comparing them to the HMM condition. This suggests that HMMs are not very effective at modeling expressive *f0* at frame-level. A possible reason for this might be the focus on frame-level modeling combined with a lack of understanding of proper supra-segmental contexts [15][16]. These models tend to average different effects, causing the generated contour to be neutral and similar to the natural low frequencies. But HMM generated *f0* is not completely similar to the lower scales, as the results from experiment 2 indicate. As for the high frequencies, we observe that they are mostly noise and do not contribute much to the naturalness of synthesized speech. This might explain the lack of improvements when modeling them at frame-level with HMMs, while using supra-segmental approaches to the other scales [4].

As future work, we propose to explore the CWT middle frequencies to model *f0* at a supra-segmental level.

Acknowledgements: This work has been conducted with the support of the Swiss NSF under grant CRSII2 141903: Spoken Interaction with Interpretation in Switzerland (SIWIS).

[2] Speech samples for all experiments can be found here: http://homepages.inf.ed.ac.uk/s1250520/samples/interspeech15.html

9. References

[1] M. H. Farouk, *Application of Wavelets in Speech Processing*. Springer, 2014.

[2] M. Vainio, A. Suni, D. Aalto *et al.*, "Continuous wavelet transform for analysis of speech prosody," *TRASP 2013-Tools and Resources for the Analysys of Speech Prosody, An Interspeech 2013 satellite event, August 30, 2013, Laboratoire Parole et Language, Aix-en-Provence, France, Proceedings*, 2013.

[3] A. S. Suni, D. Aalto, T. Raitio, P. Alku, M. Vainio *et al.*, "Wavelets for intonation modeling in hmm speech synthesis," in *8th ISCA Workshop on Speech Synthesis, Proceedings, Barcelona, August 31-September 2, 2013*, 2013.

[4] M. S. Ribeiro and R. A. J. Clark, "A multi-level representation of f0 using the continuous wavelet transform and the discrete cosine transform," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2015. ICASSP 2015.*, 2015.

[5] G. Sanchez, H. Silen, J. Nurminen, and M. Gabbouj, "Hierarchical modeling of f0 contours for voice conversion," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[6] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[7] C. Mayo, R. A. Clark, and S. King, "Multidimensional scaling of listener responses to synthetic speech," 2005.

[8] I. Recommendation, "1534-1: Method for the subjective assessment of intermediate quality level of coding systems," *International Telecommunication Union*, 2003.

[9] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[10] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the perception of prosodic prominence," *Laboratory Phonology*, vol. 1, no. 2, pp. 425–452, 2010.

[11] N. Braunschweiler, M. J. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings." in *INTERSPEECH*, 2010, pp. 2222–2225.

[12] N. Braunschweiler and S. Buchholz, "Automatic sentence selection from speech corpora including diverse speech for improved hmm-tts synthesis quality." in *INTERSPEECH*, 2011, pp. 1821–1824.

[13] T. Merritt and S. King, "Investigating the shortcomings of hmm synthesis," in *8th ISCA Workshop on Speech Synthesis (SSW8)*. Citeseer, 2013, pp. 185–190.

[14] G. E. Henter, T. Merritt, M. Shannon, C. Mayo, and S. King, "Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech." in *to appear*, 2014.

[15] O. Watts, J. Yamagishi, and S. King, "The role of higher-level linguistic features in hmm-based speech synthesis," 2010.

[16] M. Cernak, P. Motlicek, and P. N. Garner, "On the (un) importance of the contextual factors in hmm-based speech synthesis and coding," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8140–8143.