A fast algorithm for speech polarity detection using long-term linear prediction

Abhiram B.
Dept. of Electrical Engg.
Indian Institute of Science
Bangalore, India
Email: abhiram@ee.iisc.ernet.in

Prathosh A. P.
Dept. of Electrical Engg.
Indian Institute of Science
Bangalore, India
Email: prathoshap@ee.iisc.ernet.in

Ramakrishnan A. G.
Dept. of Electrical Engg.
Indian Institute of Science
Bangalore, India
Email: ramkiag@ee.iisc.ernet.in

***Abstract*—Speech polarity detection is a crucial first step in many speech processing techniques. In this paper, an algorithm is proposed that improvises the existing technique using the skewness of the voice source (VS) signal. Here, the integrated linear prediction residual (ILPR) is used as the VS estimate, which is obtained using linear prediction on long-term frames of the low-pass filtered speech signal. This excludes the unvoiced regions from analysis and also reduces the computation. Further, a modified skewness measure is proposed for decision, which also considers the magnitude of the skewness of the ILPR along with its sign. With the detection error rate (DER) as the performance metric, the algorithm is tested on 8 large databases and its performance (DER=0.20%) is found to be comparable to that of the best technique (DER=0.06%) on both clean and noisy speech. Further, the proposed method is found to be ten times faster than the best technique.**

***Index Terms*—speech polarity detection; voice source; integrated linear prediction residual; long-term linear prediction; weighted skewness**

I. Introduction

A. Motivation and importance

During the production of voiced speech, due to the nature of vocal fold vibrations, a prominent negative peak is seen in the voice source pulse (VS) waveform at the glottal closure instants (GCIs) [1]. The polarity of the speech signal is said to be positive if the prominent peak in the estimated VS signal is in its negative going half, akin to the models of the VS pulse [1], [2] and negative otherwise. Often, speech signals possess positive polarity. However, polarity reversal may be caused due to the reversal of electrical connections in the recording equipment.

There are a number of speech processing techniques where speech polarity detection (SPD) plays a crucial role. For example, in a recently proposed algorithm [3], GCIs are detected using the half-wave rectified VS that retains only the negative half. If the polarity is reversed and no effort is made to correct it beforehand, the wrong half of the VS will be retained, thus drastically affecting the subsequent analysis. Other GCI detection algorithms like zero-frequency-resonator [4] are also affected by polarity reversal. Proper GCI detection is important for algorithms, which perform pitch modification pitch synchronously [5]. Also, as pointed out in [6], [7], SPD plays a role in a unit concatenation-based text-to-speech system. Further, in recognition systems which use phase-based features, polarity detection plays an important role. For example, the speaker recognition systems proposed in [8], [9] use the discrete cosine transform coefficients of the VS as features. If training and test data come from microphones with opposite polarities, it may result in poor performance of the recognition system. The same can be said for speech recognition systems using phase-based features, like the one proposed in [10].

Thus, there is a need for accurate and robust estimation of speech polarity from the speech signal. Moreover, in data-driven approaches employing a huge amount of data, the data may originate from a variety of microphones, and the speech processing system must handle them in almost real-time. Since polarity detection is usually the very first step in an analysis procedure, it is preferable that it is computationally simple.

B. Previous work

The key approaches in the literature for SPD are summarized as follows: (a) gradient of spurious glottal waveforms method (GSGW) [11] uses the fact that the peaks in the derivative of the VS should be close to the GCIs. (b) phase cut method (PC) [12] is devised based on the observation that the first two harmonics are in phase near the GCIs. (c) relative phase shift method (RPS) [12] is based on PC but uses higher harmonics. (d) oscillating moments-based polarity detection method (OMPD) [13] uses phase shifts in the odd and even ordered statistical moments oscillating at the local fundamental frequency. (e) residual skewness method (RESKEW) [7] is based on the observation that the distribution of sample values of estimated VS will have a negative skew and that of the traditional LP residual (LPR) will have a positive skew. Accordingly, three decision rules have been presented, using skewness of only the estimated VS (RESKEW-glot), the traditional LPR (RESKEW-res), and the difference of the skewness of the VS and traditional LPR (RESKEW).

The GSGW, PC, RPS and OMPD methods require F0 and voicing decisions and are slower than the self-sufficient RESKEW method, which has been shown to have a lesser computational load [7]. Further, RESKEW has been shown to be more robust in the presence of additive noise. However, we note that it has some disadvantages such as (i) No effort is

made to avoid unnecessary analysis on the unvoiced regions which have no relevance to polarity detection, (ii) short-term analysis is carried out which imposes large computational load and (iii) analysis is carried out on both VS and traditional LPR to yield the best performance which again increases the computation.

In this paper, an SPD algorithm is proposed based on the skewness of the integrated long-term linear prediction residual. This reduces the computational complexity by eliminating the aforementioned short-comings of RESKEW without compromising much on the performance.

II. Proposed method

A. Pre-processing

1) Obtaining the ILPR: An estimate of the VS: Since the polarity is a feature that resides in the VS, it is inevitable to estimate the VS (or its variants such as LPR) from the speech signal. In this study, we use the integrated linear prediction residual (ILPR) which approximates the VS as noted in [3]. ILPR is obtained by inverse filtering the non-pre-emphasized speech signal by the vocal tract filter whose co-efficients are obtained by applying LP on the pre-emphasized speech signal. It is noteworthy that, here, pre-emphasis is realized as a first difference operation unlike RESKEW [7] where it is realized as a high-pass elliptic filter, where the choice of cut-off frequency is critical in determining the performance.

2) Low-pass filtering and selection of frame size: Normally, LP is carried out over frames of size 20-40 ms [14]. This is because, a frame too short may not include even a single pitch cycle whereas a frame too long may include two or more phones with different formant structures. Ideally, for the purpose of SPD, it is enough if the VS (or its estimate) is examined over one 'good' frame, since polarity does not change within an utterance. However, selection of this 'good' frame calls for pre-processing to determine the regions of the speech signal where an error-free estimate (computed over voiced regions, devoid of uncancelled formants, adhered to the all-pole model, etc.) of VS is obtained. The selection of this kind of a frame is a challenging task by itself, which is often computationally involved. To avoid this, the algorithms listed in Section I-B employ short-term analysis on the entire speech signal using the usual framesize of 20-40 ms with a shift of 5-10 ms. Subsequently, a majority decision is made from the polarity detected on each frame or a decision is made on the VS estimate of the entire speech signal obtained by concatenating the VS estimates of short-term frames by overlap and add method.

To lower the computational effort, long-term frames of the order of 200 ms may be used, provided the ILPR estimated on these frames preserves the polarity information. This may not be the case if the usual speech signal is used due to improper formant-cancellation by inverse filtering. To avoid improper formant-cancellation over long-term frames, we use the low-pass filtered (LPF) speech signal to obtain the ILPR. From the well-known Peterson and Barney experiment [15], it can be seen that F2 and higher formants are above 800 Hz for most vowels. Hence, the cut-off frequency of the low-pass filter is chosen to be 800 Hz, thereby attenuating the effect of F2 and higher formants. This does not alter the gross shape of the VS estimate as noted in [11]. However, lowering the cut-off further may attenuate the glottal formant also, thereby destroying the gross shape of the VS, which is needed for SPD. Here, the low-pass filtering is implemented in the frequency domain, with a flat response in the pass-band, a raised cosine-shaped roll-off in the transition-band, and zero in the stop-band. This attenuates the high frequency components and thereby reduces the energy in the unvoiced regions also. A simple energy threshold would now allow to discard the frames which are predominantly unvoiced.

Note that, even after low-pass filtering, the effect of F1 is present, which is to be cancelled to get an accurate estimate of the VS. For this, ILPR is obtained on the long-term frames of the LPF speech signal. The order q used for LP analysis is the usual order, that is, $q = f_s(kHz)+2$, where f_s is the sampling frequency. A lower order may lead to estimation of sharp (low-bandwidth) spectral peaks, whereas the actual spectral envelope may be wider, resulting in poor inverse filtering. An example of a 200 ms segment of a speech signal sampled at 16 kHz consisting of two vowels is shown in Fig. 1(a). The ILPR estimated from the LPF version of the signal using LP orders 4 and 18 are shown in Figs. 1(b) and 1(c), respectively. It may be seen that the ILPR obtained using $q = 4$ resembles the speech signal in Fig. 1(a), whereas the ILPR obtained using $q = 18$ has the expected gross shape of the VS, retaining the prominent negative peaks.

The spectrum of the LPF speech signal (filled line) and the envelopes obtained from LP using $q = 4$ (dashed line) and $q = 18$ (dotted line) are also shown in Fig. 2. It can be observed that one sharp (low-bandwidth) peak is estimated using $q = 4$ leading to improper inverse filtering but with $q = 18$, the spectral envelope is estimated relatively better resulting in better inverse filtering. The resulting ILPR is again low-pass filtered with a cut-off frequency of 800 Hz to remove the high frequency components resulting from inverse filtering.

B. Decision rule: The Weighted Skewness measure

The decision rule is based on the skewness of the estimated VS as proposed in RESKEW-glot [7]. Thus, polarity is positive if skewness of the distribution of the VS estimate values is negative and vice-versa. This is because, if the polarity is positive, there will be lesser number of samples in the negative-half of the VS estimate compared to the positive-half and the magnitude of the negative peak will be higher than that of the positive peak in every cycle. This results in a distribution having a longer tail to the left quantified by a negative skew. The scenario is reversed if there is a polarity reversal.

As already mentioned, RESKEW-glot [7] uses the skew of the VS estimate of the entire utterance. However, long-term analysis gives the luxury to take framewise decisions, and hence the decision rule here is modified to include a weight-factor with the skewness measure as proposed in the

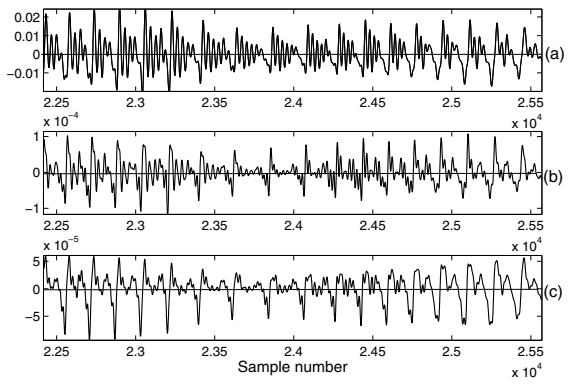

Fig. 1. Illustration of estimation of ILPR on the low-pass filtered speech signal. (a) A 200 ms speech segment consisting an /au/ followed by an /e/. (b) ILPR estimated from the low pass component of the speech signal shown in (a) with $q = 4$. (c) ILPR estimated from the low pass component of the speech signal shown in (a) with $q = 18$. The gross-shape of the VS is well preserved in the signal shown in (c) but not in (b).

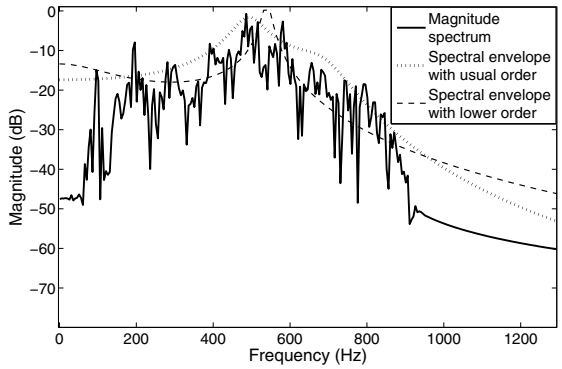

Fig. 2. Illustration of the effect of LP order on the estimation of the ILPR. Filled line: spectrum of the LPF speech signal shown in Fig. 1(a). Dotted line: spectral envelope with $q = 18$. Dashed line: spectral envelope with $q = 4$. It is seen that the envelope is relatively better estimated using $q = 18$.

next section. Since the skewness value gives a measure of the asymmetry of the distribution, more negative the skew, more likely is the polarity to be positive. Also, more the number of frames which have negative skew, more likely is the polarity to be positive. Thus, the sum of the frame-wise negative skew values is weighted by the number of frames having negative skew and the result is compared with the sum of the frame-wise positive skew values weighted by the number of frames having positive skew, and the polarity is decided to be positive if the former has the larger magnitude.

C. The proposed algorithm

This section summarizes the algorithm, which we term as the long-term weighted skew (LT-WSKEW) algorithm.

1) Low-pass filter the input speech signal $s(n)$ with a cut-off frequency $f_c = 800\ Hz$ to get $s_f(n)$.
2) Divide $s_f(n)$ into non-overlapping frames of length 200 ms. For each frame, perform steps 3 to 5.
3) If the energy of a frame normalized by the average energy over all frames is less than a threshold (0.75 used here), reject it as a possible unvoiced frame and go to the next frame.
4) Pre-emphasize $s_f(n)^i$ to get $s'_f(n)^i$, where i is the frame index.
5) Inverse filter $s_f(n)^i$ using the LP co-efficients estimated from $s'_f(n)^i$ with LP order $q = f_s(kHz) + 2$ to get the ILPR, $e(n)^i$.
6) Compute S^i, the skewness of $e(n)^i$, for all frames.
7) Compute a Weighted Skewness (WS) measure

$$WS = N_p * \sum_{\{x:S^x>0\}} S^x + N_n * \sum_{\{y:S^y<0\}} S^y \qquad (1)$$

where N_p and N_n are the number of frames for which S^i is positive and negative respectively; S^x and S^y are the skewness values for the frames with positive and negative skews respectively.
8) Polarity of $s(n)$ is deemed to be positive if WS is negative and vice-versa.

III. Experimental procedure and results

A. Databases and performance measures

The algorithm is tested on eight large corpora for a fair comparison of its performance with that of [7]. Among them, seven are from the CMU ARCTIC database [16] created for the purpose of TTS: AWB (male), BDL (male), CLB (female), JMK (male), KSP (male), RMS (male) and SLT (female). The last one is a German emotional speech (7 emotions, 10 speakers, 5 male and 5 female) corpus called the Berlin corpus [17]. There are a total of 8446 files over the 8 corpora totalling to about six hours of speech data. Each file is presented as a trial and the accuracy of the LT-WSKEW algorithm is measured by the detection error rate (DER), which is the ratio of the number of wrong detections to the total number of trials. The ground truth is determined by visually examining the short-term VS estimates of several utterances in a given database, which is the standard method followed. The computational load is measured in terms of Relative Computation Time (RCT), which is the ratio of the time taken for algorithm execution to the duration of the utterance [7]. Experiments are conducted both in clean conditions and in the presence of the additive babble noise at different signal-to-noise ratios (SNRs).

B. Effect of cut-off frequency

Fig. 3 shows the performance variation of LT-WSKEW with the cut-off frequency of the low-pass filter on the data from CMU ARCTIC speaker AWB (with 1138 trials). It is clear that the performance is most affected when the cut-off frequency decreases below 300 Hz, showing that the frequency components till 300 Hz, which include the glottal formant and the fundamental, contain crucial polarity information. Components with frequency higher than 800 Hz do not affect the performance at all, showing that the long-term LP analysis is robust and does not allow F2 and higher formants to affect

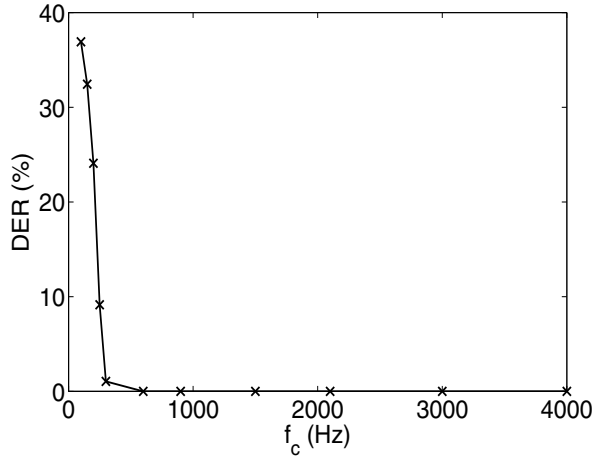

Fig. 3. Performance variation with cut-off frequency f_c of the low-pass filter

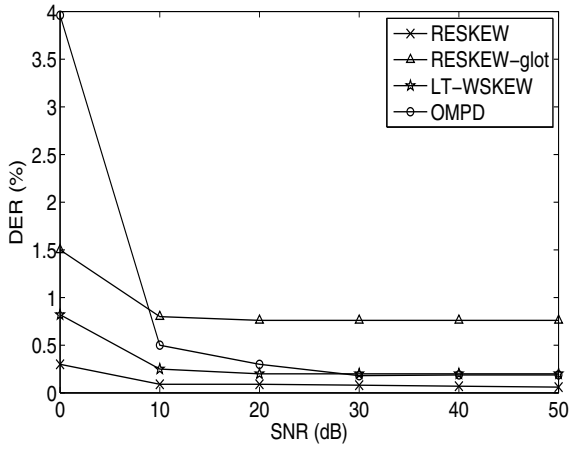

Fig. 4. Performance comparison of LT-WSKEW with three other polarity detection algorithms on speech with additive babble noise

the polarity information embedded in the skewness of the ILPR.

C. Results on clean speech

The proposed algorithm gives a DER of 0.2% (17 wrong detections), which is almost equal to the second best method among those compared in [7], OMPD, which gives a DER of 0.187% (16 wrong detections). But note that OMPD requires voicing and F0 decisions and short-term analysis and hence is computationally more expensive than LT-WSKEW. The best algorithm is RESKEW with a DER of 0.0592% (5 wrong detections), but note that LT-WSKEW does not use skewness of the traditional LPR which RESKEW uses. RESKEW-glot which uses only skewness of the VS gives a DER of 0.76% (64 wrong detections). The improvement shown by LT-WSKEW compared to RESKEW-glot can be attributed to the low-pass filtering and the different pre-emphasis filter used (first difference in the case of LT-WSKEW and elliptic filter in the case of RESKEW-glot) and also to the modified skewness measure used. It has been noted that all except one of the wrong detections of LT-WSKEW arise from the Berlin speech corpus. This may be due to the large non-stationary nature of the emotional speech which leads to incorrect LP estimates. It is noteworthy that DER increases to 0.56% if a majority decision is taken by just comparing the number of frames having negative and positive skews instead of weighted skew.

D. Results on noisy speech

Results with additive babble noise taken from the Noisex-92 database [18] are shown in Fig. 4. At 0 dB SNR, LT-WSKEW performs better than all other algorithms except RESKEW. At higher SNRs, LT-WSKEW performs atleast as well as the second best algorithm, whereas RESKEW is always the best algorithm. The good performance of LT-WSKEW can be attributed to the fact that it uses the skewness of the VS, whose discontinuities are not affected much in the presence of additive noise [7].

E. Computational complexity

RESKEW has been shown to have the least RCT in [7]. Since RESKEW-glot is a component of RESKEW, it is observed to have a lesser RCT than RESKEW and hence is the fastest of the existing algorithms. Hence, here RCT is computed for RESKEW-glot and LT-WSKEW for comparison using a MATLAB implementation on an INTEL i5-2410M, 2.30 GHz processor. Fig. 5 shows histograms of RCTs of LT-WSKEW and RESKEW-glot. One can see a large dynamic range of the RCT for RESKEW-glot spanning between 10 and 50, which is greater than that of LT-WSKEW by a factor of 10. Also, RESKEW-glot and LT-WSKEW have average RCTs of 21.7% and 2.0%, respectively. Thus, LT-WSKEW is, on an average, 10 times faster than the fastest existing algorithm.

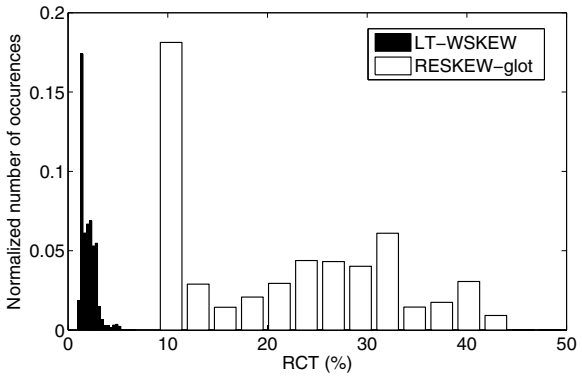

Fig. 5. RCT histograms of LT-WSKEW and RESKEW-glot

The average RCTs of the component blocks involved in the LT-WSKEW and RESKEW-glot algorithms are shown in Table I. It is apparent that the average RCT of the inverse filtering block in RESKEW-glot (8.05%) is more than ten times that of the same block in LT-WSKEW (0.55%). The low-pass filtering block in LT-WSKEW is an added load, but the average RCT is only 0.11%, which is almost negligible when compared to that

TABLE I
AVERAGE RCT OF THE COMPONENT BLOCKS OF LT-WSKEW AND RESKEW-GLOT

Algorithm	Average RCT (%)		
	Low-pass filtering	Inverse filtering	Decision
LT-WSKEW	0.11	**0.55**	0.16
RESKEW-glot	–	**8.05**	0.21

of the inverse filtering block in RESKEW-glot. The RCTs of the decision block, involving skew computation and decision logic, is comparable between the two algorithms. Even though RESKEW-glot involves a single skewness computation over the entire utterance, the decision block in LT-WSKEW has a lesser average RCT because the unvoiced frames are rejected in LT-WSKEW but retained in RESKEW-glot.

IV. CONCLUSION

A simple and fast algorithm is proposed for speech polarity detection using the weighted skewness of long-term LP residual. The ILPR is used as the VS estimate, and long-term estimates of the ILPR are obtained using LP analysis on the LPF speech signal, which drastically reduces the computational load yet yielding comparable results. The low frequency components of the VS, which include the glottal formant and the fundamental are found to contain significant polarity information. The performance of the proposed algorithm is found to be comparable to that of the existing algorithms on clean speech as well as speech with additive babble noise. Further, the algorithm is 10 times faster than the fastest of the existing methods which has practical significance.

ACKNOWLEDGMENT

The authors wholeheartedly thank Dr. T.V. Ananthapadmanabha for introducing them to the ILPR and explaining it in detail.

REFERENCES

[1] T.V. Ananthapadmanabha, "Acoustic analysis of voice source dynamics," *STL-QPSR*, vol. 25, no. 2-3, pp. 1–24, 1984.

[2] G. Fant, J. Liljencrants, and Q. Lin, "A four parameter model of glottal flow," *STL-QPSR*, vol. 4, pp. 1–13, 1985.

[3] A.P. Prathosh, T.V. Ananthapadmanabha, and A.G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 21, no. 12, pp. 2471–80, 2013.

[4] K.S.R Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 16, no. 8, pp. 1602–13, 2008.

[5] R. Muralishankar, A. G. Ramakrishnan, and P. Prathibha, "Modification of pitch using DCT in the source domain," *Speech Communication*, vol. 42, no. 2, pp. 143–154, 2004.

[6] A.J. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP*, 1996, vol. 1, pp. 373–376.

[7] T. Drugman, "Residual excitation skewness for automatic speech polarity detection," *Signal Processing Letters, IEEE*, vol. 20, no. 4, pp. 387–390, 2013.

[8] A. G. Ramakrishnan, B. Abhiram, and S.R.M. Prasanna, "A characterization of the voice source using pitch synchronous discrete cosine transform for speaker information," *unpublished*.

[9] Rohan Kumar Das, B. Abhiram, S.R.M. Prasanna, and A. G. Ramakrishnan, "Combining source and system information for limited data speaker verification," *unpublished*.

[10] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *ICASSP*, 2001, vol. 1, pp. 133–136.

[11] W. Ding and N. Campbell, "Determining polarity of speech signals based on gradient of spurious glottal waveforms," in *ICASSP*, 1998, vol. 2, pp. 857–860.

[12] I. Saratxaga, D. Erro, I. Hernez, I. Sainz, and E. Navas, "Use of harmonic phase information for polarity detection in speech signals," in *Interspeech*, 2009, pp. 1075–78.

[13] T. Drugman and T. Dutoit, "Detecting speech polarity with high-order statistics," *Cognitive Computation Journal, special issue of NOLISP11, doi: 10.1007/s12559-012-9167-y*, 2012.

[14] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[15] G.E. Peterson and H.L. Barney, "Control methods used in a study of the vowels," *Journal of the Acoustical Society of America*, vol. 24, pp. 175 – 184, 1952.

[16] J. Kominek and A. Black, "The CMU Arctic speech databases," *SSW5*, pp. 223–224, 2004.

[17] F. Burkhardt, A. Paseschke, and M. Rolfes, "A database of German emotional speech," in *Interspeech*, 2005, pp. 1517–20.

[18] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II NOISEX92, a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.