BLIND SPATIAL SUBTRACTION ARRAY WITH INDEPENDENT COMPONENT ANALYSIS FOR HANDS-FREE SPEECH RECOGNITION

Yu Takahashi, Tomoya Takatani, Hiroshi Saruwatari and Kiyohiro Shikano

{yuu-t, tomoya-t, sawatari, shikano}@is.naist.jp
Nara Institue of Science and Technology, Nara, 630-0192, JAPAN

ABSTRACT

In this paper, we propose a new blind spatial subtraction array (BSSA) which contains an accurate noise estimator based on independent component analysis (ICA) to realize a noise-robust hands-free speech recognition. First, a preliminary experiment suggests that the conventional ICA is proficient in the noise estimation rather than the direct speech estimation in real environments, where the target speech can be approximated to a point source but real noises are often not point sources. Secondly, based on the above-mentioned findings, we propose a new noise reduction method which is implemented in subtracting the power spectrum of the estimated noise by ICA from the power spectrum of noise-contaminated observations. This architecture provides us with a noise-estimation-error robust speech enhancement which is well applicable to the speech recognition. Finally, the effectiveness of the proposed BSSA is shown in the speech recognition experiment.

1. INTRODUCTION

A hands-free speech recognition system is essential for realizing an intuitive, unconstrained, and stress-free human-machine interface. In this system, however, it is difficult to achieve a high recognition accuracy because noise and the reverberation always deteriorate a target speech quality. One approach to address the problem is to separate the observed signals into each original signal by blind source separation (BSS) technique. BSS is the approach to estimate the original sources using only information of the observed signal in each microphone. Basically, BSS is classified as an unsupervised filtering technique, and does not require any supervisions on directions-of-arrival (DOAs) and target-speech pause where only noise exists.

Recently, various methods of BSS based on independent component analysis (ICA) [1] have been presented on acoustic-sound separation [2, 3, 4, 5]. Indeed the conventional ICA could work especially in speech-speech (or point sources) mixing, but such a mixing condition is very rare and not realistic; real noises are often widely-spread sources. In this paper, first, we show a result of preliminary experiment which tells that ICA is proficient in the noise estimation rather than the speech estimation when noise is not a point source. Based on the above-mentioned fact, then we propose a new blind spatial subtraction array (BSSA) with an ICA-based noise estimation, which is achieved by subtracting the power spectrum of the estimated noise via ICA from the power spectrum of the noisy observations. This "power-spectrum-domain subtraction" procedure provides a better noise reduction than the conventional ICA with a estimation-error robustness. Finally, the real-recording-based simulations are conducted, and we can indicate that the proposed BSSA outperforms the conventional methods on the improvements in noise reduction and speech recognition.

2. IS ICA PROFICIENT IN TARGET-SPEECH ESTIMATION OR NOISE ESTIMATION?

Many previous researches on BSS provided evidences in that the conventional ICA could work in source separation, especially for the special case of speech-speech mixing. However, such a sound mixing is not realistic under common acoustic conditions; indeed the following scenario and problem are likely to arise (see Fig. 1).

- The target sound is user's speech, which can be approximately regarded as a *point source*. In addition, the user locates themselves relatively *close to the microphone array* (e.g., 1 m apart), and consequently the accompanying reflection and reverberation components are small.
- As for the noise, we are often confronted with interference sounds which are *not point sources* but widely-spread sources. Also the noise is usually far from the array and heavily reverberant.
- From the above-mentioned scenario, it is expected that the conventional ICA can suppress the user's speech signal to pick up the noise source, but the ICA is very weak in picking up target speech itself via suppression of the far-located widely-spread noise. This is due to the fact that ICA with the small number of sensors and filter taps often provides only directional nulls against the undesired source signals [5].

Figure 2 illustrates a real separation result (noise reduction rate (NRR) [4] defined in Sect. 4.2) of the conventional ICA obtained in a preliminary experiment, where the noise's NRR is calculated in the case that the cleaner noise is regarded as the target signal. The experimental conditions are the same as those in Sect. 4.1 except for the number of microphones (=2). This result gives us an unfortunate conclusion that ICA is *not* proficient in target-speech estimation. However, this also implies that we can still use ICA as an accurate noise estimator even under reverberant conditions.

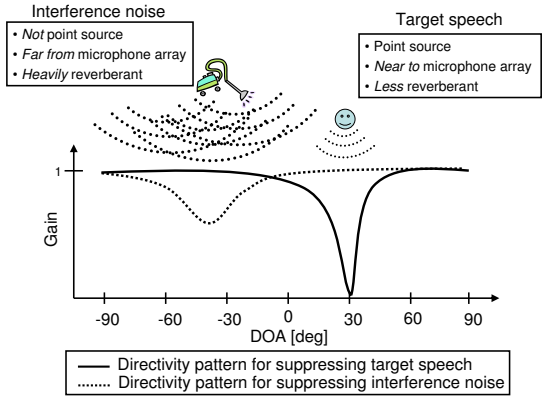

Figure 1: *Directivity pattern which is shaped by ICA.*

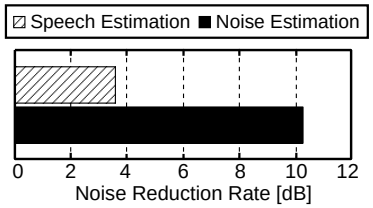

Figure 2: *NRR-based performance of conventional ICA in environment shown in Fig. 4.*

3. PROPOSED METHOD

3.1. Motivation and Strategy

The consideration described in the previous section motivates us to propose a new speech-enhancement strategy, i.e., BSSA. The proposed method consists of a delay-and-sum array (DS)[6] based primary path and a reference path for the ICA-based noise estimation (see Fig. 3). The estimated noise component by ICA is efficiently subtracted from the primary path in the power-spectrum domain without phase information. This procedure can yield a better target-speech enhancement than the simple ICA, even with a benefit of estimation-error robustness in the speech recognition application. The detailed signal processing is shown below.

3.2. Partial Speech Enhancement in Primary Path

First, the short-time analysis of observed signals is conducted by a frame-by-frame discrete Fourier transform (DFT). By plotting the spectral values in a frequency bin for each microphone input frame by frame, we consider these values as a time series. Hereafter, we designate the time series as

$$\boldsymbol{X}(f,\tau) = [X_1(f,\tau), \cdots, X_J(f,\tau)]^{\mathrm{T}}, \tag{1}$$

where f is the frequency bin and τ is the frame number. Also, $\boldsymbol{X}(f,\tau)$ can be rewritten as

$$\boldsymbol{X}(f,\tau) = \boldsymbol{A}(f)\left(\boldsymbol{S}(f,\tau) + \boldsymbol{N}(f,\tau)\right), \tag{2}$$

$$\boldsymbol{S}(f,\tau) = [\underbrace{0,\cdots,0}_{U-1}, S_U(f,\tau), \underbrace{0,\cdots,0}_{K-U}]^{\mathrm{T}}, \tag{3}$$

$$\boldsymbol{N}(f,\tau) = [N_1(f,\tau),..., N_{U-1}(f,\tau), 0, N_{U+1}(f,\tau),..., N_K(f,\tau)]^{\mathrm{T}}, \tag{4}$$

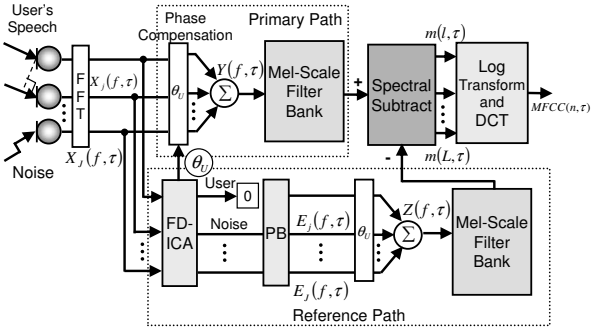

Figure 3: *Diagram of proposed BSSA.*

where $\boldsymbol{A}(f)$ is a mixing matrix, $\boldsymbol{S}(f,\tau)$ is a target speech signal vector, $\boldsymbol{N}(f,\tau)$ is a noise signal vector, U expresses the target speech number, and K is the number of sound sources.
Next, the target speech signal is partly enhanced in advance by DS. This procedure can be given as

$$\begin{aligned} Y(f,\tau) &= \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{X}(f,\tau) \\ &= \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{A}(f)\boldsymbol{S}(f,\tau) \\ &\quad + \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{A}(f)\boldsymbol{N}(f,\tau), \end{aligned} \tag{5}$$

$$\boldsymbol{W}_{\mathrm{DS}}(f) = [W_1^{(\mathrm{DS})}(f), \ldots, W_J^{(\mathrm{DS})}(f)]^{\mathrm{T}}, \tag{6}$$

$$W_j^{(\mathrm{DS})}(f) = \frac{1}{J}\exp\left(-i2\pi(f/M)f_s d_j \sin\theta_U/c\right), \tag{7}$$

where $Y(f,\tau)$ is a primary-path output which a slightly enhances target speech, $\boldsymbol{W}_{\mathrm{DS}}(f)$ is a filter coefficient vector of DS [6], M is the DFT size, f_s is sampling frequency, d_j is a microphone position, and c is sound velocity. Besides, θ_U is the estimated DOA of the target speech which is given by ICA part in Sect. 3.3. In Eq. (5), the second term in the right-hand side expresses the remaining noise in the output of the primary path.

3.3. ICA-Based Noise Estimation in Reference Path

The proposed BSSA provides ICA-based noise estimation. In ICA, we perform signal separation using the complex valued unmixing matrix $\boldsymbol{W}_{\mathrm{ICA}}(f)$, so that the output signals $\boldsymbol{O}(f,\tau) = [O_1(f,\tau),\ldots,O_J(f,\tau)]^{\mathrm{T}}$ become mutually independent; this procedure can be represented by

$$\boldsymbol{O}(f,\tau) = \boldsymbol{W}(f)\boldsymbol{X}(f,\tau), \tag{8}$$

$$\boldsymbol{W}(f) = \boldsymbol{P}(f)\boldsymbol{W}_{\mathrm{ICA}}(f), \tag{9}$$

where $\boldsymbol{P}(f)$ is a permutation matrix and $\boldsymbol{W}(f)$ is a new unmixing matrix which resolves the permutation problem. The permutation matrix $\boldsymbol{P}(f)$ is determined by looking at null directions in the directivity pattern which is shaped by $\boldsymbol{W}_{\mathrm{ICA}}(f)$ [4], so that the U-th output $O_U(f,\tau)$ is set to the target speech signal. At the same time, we can estimate DOAs, and we designate DOA of the target speech signal as θ_U. The optimal $\boldsymbol{W}_{\mathrm{ICA}}(f)$ is obtained by the following iterative updating equation [2]:

$$\begin{aligned} \boldsymbol{W}_{\mathrm{ICA}}^{[i+1]}(f) &= \mu\left[\boldsymbol{I} - \langle\boldsymbol{\Phi}\left(\boldsymbol{O}(f,\tau)\right)\boldsymbol{O}^{\mathrm{H}}(f,\tau)\rangle_\tau\right]\boldsymbol{W}_{\mathrm{ICA}}^{[i]}(f) \\ &\quad + \boldsymbol{W}_{\mathrm{ICA}}^{[i]}(f), \end{aligned} \tag{10}$$

where μ is the step-size parameter, $[i]$ is used to express the value of the i-th step in the iterations, and $\boldsymbol{I}$ is an identity matrix. Besides, $\langle\cdot\rangle_\tau$ denotes a time-averaging operator, $\boldsymbol{M}^{\mathrm{H}}$ denotes

hermitian transpose of matrix $\boldsymbol{M}$, and $\boldsymbol{\Phi(\cdot)}$ is the appropriate nonlinear vector function [4].In the reference path, target signal is not required because we want to estimate only the noise component. Accordingly we remove the separated speech component $O_U(f,\tau)$ from ICA outputs $\boldsymbol{O}(f,\tau)$, and construct the following "noise-only vector", $\boldsymbol{Q}(f,\tau)$;

$$\boldsymbol{Q}(f,\tau) = [O_1(f,\tau),...,O_{U-1}(f,\tau),0,O_{U+1}(f,\tau),...,O_J(f,\tau)]^{\mathrm{T}}. \quad (11)$$

Next, we apply the projection back (PB) [3] method to remove the ambiguity of amplitude. This procedure can be represented as

$$\boldsymbol{E}(f,\tau) = \boldsymbol{W}^{+}(f)\boldsymbol{Q}(f,\tau), \quad (12)$$

where $\boldsymbol{M}^{+}$ denotes Moore-Penrose pseudo inverse matrix of $\boldsymbol{M}$. Here, $\boldsymbol{Q}(f,\tau)$ is composed of only noise components. Therefore, $\boldsymbol{E}(f,\tau)$ is a good estimation of the received noise signals at the microphone positions;

$$\boldsymbol{E}(f,\tau) \simeq \boldsymbol{A}(f)\boldsymbol{N}(f,\tau). \quad (13)$$

Finally, we obtain the estimated noise signal $Z(f,\tau)$ by performing DS as follows:

$$Z(f,\tau) = \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{E}(f,\tau) \simeq \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{A}(f)\boldsymbol{N}(f,\tau). \quad (14)$$

Equation (14) is expected to be equal to the noise term of Eq. (5) in the primary path. Of course, Eq. (14) contains estimation errors to some extent. Even though the level of the noise estimation error is not negligible, we can still enhance the target speech via over-subtraction[8] in the power-spectrum domain. Note that $Z(f,\tau)$ is the function of the frame number τ, unlike the constant noise prototype estimated in the traditional spectral subtraction method [8]. Therefore, the proposed BSSA can deal with *non-stationary* noise.

3.4. Noise Reduction Processing

The proposed BSSA includes mel-scale filter bank analysis, and directly outputs mel-frequency cepstrum coefficient (MFCC) [7]. The triangular window $W_{\mathrm{mel}}(k;l)$ $(l = 1,\cdots,L)$ to perform mel-scale filter bank analysis is designated as

$$W_{\mathrm{mel}}(f;l) = \begin{cases} \dfrac{f - f_{\mathrm{lo}}(l)}{f_{\mathrm{c}}(l) - f_{\mathrm{lo}}(l)} & \big(f_{\mathrm{lo}}(l) \le f \le f_{\mathrm{c}}(l)\big), \\ \dfrac{f_{\mathrm{hi}}(l) - f}{f_{\mathrm{hi}}(l) - f_{\mathrm{c}}(l)} & \big(f_{\mathrm{c}}(l) \le f \le f_{\mathrm{hi}}(l)\big), \end{cases} \quad (15)$$

where $f_{\mathrm{lo}}(l)$, $f_{\mathrm{c}}(l)$, and $f_{\mathrm{hi}}(l)$ are the lower, center, and higher frequency bins of each triangle window, respectively. Furthermore, L is the dimension of mel-scale filter bank. They satisfy the relation among adjacent windows as

$$f_{\mathrm{c}}(l) = f_{\mathrm{hi}}(l-1) = f_{\mathrm{lo}}(l+1). \quad (16)$$

Moreover, $f_{\mathrm{c}}(l)$ is arranged in regular intervals on mel-frequency domain. Mel-scale frequency $Mel_{f_{\mathrm{c}}(l)}$ for $f_{\mathrm{c}}(l)$ is calculated as

$$Mel_{f_{\mathrm{c}}(l)} = 2595\log_{10}\{1 + k_{\mathrm{c}}(l)f_s/(700{\cdot}M)\}. \quad (17)$$

In the proposed BSSA, noise reduction is carried out by subtracting the estimated noise power spectrum (Eq. (14)) from the enhanced target speech power spectrum (Eq. (5)) in the mel-scale filter bank domain. This procedure is defined as follows:

$$m(l,\tau) = \begin{cases} \displaystyle\sum_{f=f_{lo}(l)}^{f_{hi}(l)} W_{\mathrm{mel}}(f;l)\{|Y(f,\tau)|^2 - \beta\cdot|Z(f,\tau)|^2\}^{\frac{1}{2}} \\ \qquad (\text{ if } \ |Y(f,\tau)|^2 - \beta\cdot|Z(f,\tau)|^2 \ge 0\ ), \\ \displaystyle\sum_{f=f_{lo}(l)}^{f_{hi}(l)} W_{\mathrm{mel}}(f;l)\{\gamma\cdot|Y(f,\tau)|\} \ (\text{otherwise}), \end{cases} \quad (18)$$

where $m(l,\tau)$ is the output from the mel-scale filter bank, $Y(f,\tau)$ is the output signal from the primary path, i.e., the partially enhanced speech signal, and $Z(f,\tau)$ is the output signal from the reference path, i.e., the estimated noise signal.
The system switches in two equations depending on the conditions in Eq. (18). If the calculated noise components by ICA (Eq. (14)) are underestimated, i.e., $|Y(f,\tau)|^2 > \beta|Z(f,\tau)|^2$, the resultant output $m(l,\tau)$ corresponds to the power-spectrum-domain subtraction among primary and reference paths with the over-subtraction parameter of β. On the other hand, if the noise components are overestimated in ICA, the resultant output $m(l,\tau)$ is floored with a small positive value to avoid the negative-valued unrealistic spectrum. These *over-subtraction* and *flooring* procedures promise us an error-robust speech enhancement in the proposed BSSA rather than a simple linear subtraction. Although the nonlinear processing in Eq. (18) often generates an artificial distortion, the so called *musical noise*, it is still applicable in a speech recognition system because the speech decoder is not so sensitive to such a distortion.
Moreover, the proposed BSSA is performed in the mel-scale filter bank domain, so that transformation into MFCC can be easily performed as

$$MFCC(n,\tau) = \sqrt{\frac{2}{L}}\sum_{l=1}^{L}\log\{m(l,\tau)\}\cos\left\{\left(l-\frac{1}{2}\right)\frac{n\pi}{L}\right\}, \quad (19)$$

where n denotes the dimension of MFCC. The proposed BSSA doesn't require transformation into the time-domain waveform.

4. EXPERIMENTS AND RESULT

4.1. Experimental Setup

Figure 4 shows a layout of the reverberant room used in our experiments. We used the following 16 kHz sampled signals as test data; the original speech convoluted with the impulse responses recorded in the real environment, and added with a cleaner noise which was recored in the real environment. The cleaner noise is not a point source but consists of several non-stationary noises emitted from, e.g., a motor, air duct and nozzle. The input signal-to-noise ratio (SNR) is set to 5, 10, or 15 dB at the array. A four-element array with the interelement spacing of 2 cm is used, and DFT size is 512. Over-subtraction paramaeter β is 1.4 and flooring coefficient γ is 0.2.

4.2. Results of Noise Reduction Performance

We compared DS, the conventional ICA, and the proposed BSSA on the basis of NRR [4], which is defined as the output SNR in dB minus the input SNR in dB. In this experiment, we used 6 speakers (6 sentences) as original speech. Figure 5 shows average of the NRRs for each method. From this result, we can confirm that the NRR of the proposed BSSA overtakes those of DS and ICA by more than 4 dB. This indicates that the proposed BSSA is beneficial to realistic noise reduction applications.

Table 1: *Conditions for speech recognition*

Database	JNAS [9], 306 speakers (150 sentences / 1 speaker)
Task	20 k newspaper dictation
Acoustic model	phonetic tied mixture (PTM) [9], clean model
Number of training speakers for acoustic model	260 speakers (150 sentences / 1 speaker)
Decoder	JULIUS [9] ver 3.5.1

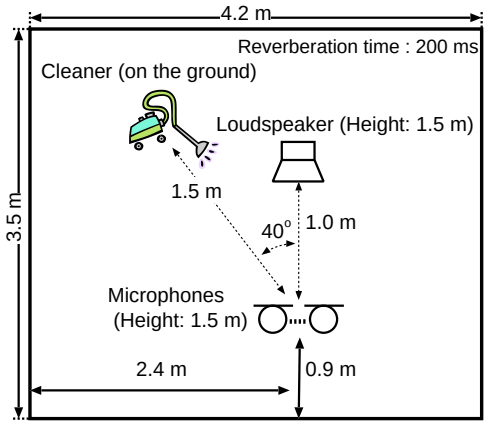

Figure 4: *Layout of reverberant room used in our experiment.*

4.3. Results of Speech Recognition Performance

We compared DS, the conventional ICA, the conventional single-channel spectral subtraction [8] cascaded with the ICA (ICA+SS), and the proposed BSSA on the basis of word accuracy scores. Table 1 shows the conditions for speech recognition, and we used 46 speakers (200 sentences) as original speech.

Figure 6 shows the word accuracy in each method. Here, "Unprocessed" refers to the result without any noise reduction processing. From this result, we can see that the word accuracy of the proposed BSSA is obviously superior to those of the conventional methods. It should be mentioned that the proposed BSSA can still outperform the simple combination of existing ICA and SS. This is a promising evidence that the proposed BSSA has an applicability to noise-robust speech recognition.

5. CONCLUSIONS

In this paper, we proposed a new BSSA which involves ICA-based noise estimation to realize a robust hands-free speech recognition in noisy environments. First, a preliminary experiment pointed out the fact that ICA is proficient in the noise estimation when noise is not a point source. Secondly, based on the above-mentioned findings, we proposed a new noise reduction strategy which is achieved by subtracting the power spectrum of the estimated noise via ICA from the power spectrum of the noisy observations. Finally, it was confirmed that the word accuracy of the proposed BSSA overtook those of DS, ICA and ICA+SS in the experiment.

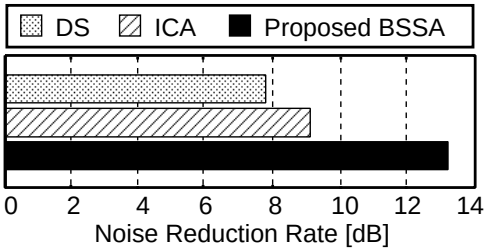

Figure 5: *Results of noise reduction rate in each method.*

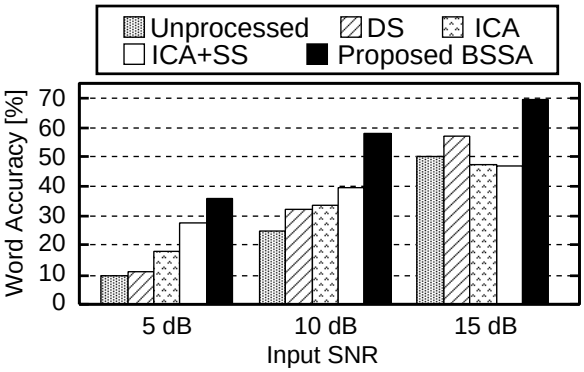

Figure 6: *Results of word accuracy in each method.*

6. ACKNOWLEDGMENT

The work was partly supported by MEXT e-Society leading project.

7. REFERENCES

[1] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol.36, pp.287–314, 1994.

[2] P. Smaragdis. "Blind separation of convoluted mixtures in the frequency domain," *Neurocomputing*, vol.22, no.1-3, pp.21–34, 1998.

[3] S. Ikeda and N. Murata, "A method of ICA in the frequency domain," *Proceedings of International Workshop on Independent Component Analysis and Blind Signal Separation*, pp.365–371, 1999. vol.20, pp.229–240, 1996.

[4] H. Saruwatari, et al., "Blind source separation combining independent component analysis and beamforming," *EURASIP J. Applied Signal Proc.*, vol.2003, no.11, pp.1135–1146, 2003.

[5] S. Araki, et al., "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Transactions on Speech and Audio Processing*, Vol.11, No.2, pp.109-116, 2003.

[6] J. L. Flanagan, et al., "Computer-steered microphone arrays for sound tranduction in large rooms," *J. Acoust. Soc. America*, vol.78, no.5, pp.1508–1518, 1985.

[7] S. B. Davis, et al., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol.ASSP-28, no.4, pp.357–366, 1982.

[8] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoustics, Speech, Signal Proc*, vol.ASSP-27, no.2, pp.113–120, 1979.

[9] A. Lee, et al., "Julius – An open source real-time large vocabulary recognition engine," *Proc. EUROSPEECH*, pp.1691–1694, 2001.